Learning with random forests

Erwan Scornet

Sciences mathématiques de Paris Centre

Thèse de doctorat

Discipline : Mathématiques

Spécialité : Statistiques

présentée par

Erwan Scornet

Apprentissage et forêts aléatoires

---

dirigée par Gérard Biau et par Jean-Philippe Vert

Au vu des rapports établis par
MM. Peter Bühlmann et Pierre Geurts

Soutenue le 30 novembre 2015 devant le jury composé de :

M. Sylvain Arlot	Université Paris-Sud	Examinateur
M. Gérard Biau	Université Paris 6	Directeur de thèse
M. Pierre Geurts	Université de Liège	Rapporteur
M. Arnaud Guyader	Université Paris 6	Examinateur
M. Jean-Philippe Vert	Mines ParisTech - Institut Curie	Directeur de thèse

Laboratoire de Statistique Théorique et Appliquée (LSTA)
Université Pierre et Marie Curie
Boîte 158, Tours 15-25, 2ème étage
4 place Jussieu
75252 Paris Cedex 05

Centre for Computational Biology (CBIO)
MINES ParisTech
60 boulevard Saint-Michel
75006 Paris

Ce n'est qu'en essayant continuellement que l'on finit par réussir. Autrement dit : plus ça rate, plus on a de chances que ça marche.

Devise Shadok

Remerciements

Avant de nous enfoncer dans les méandres des forêts aléatoires, arrêtons-nous un instant sur ce qui a permis la soutenance de cette thèse.

Je tiens tout d'abord à remercier Gérard et Jean-Philippe qui m'ont encadré et m'ont constamment apporté leur soutien et leurs précieux conseils. Ils m'ont également permis de prendre du recul sur les mathématiques, et de ne pas douter lorsqu'une difficulté venait se glisser dans des démonstrations de plusieurs pages. Ils m'ont appris à persévérer, à continuer de chercher, même lorsque les équations semblaient contre moi. Cela a été un véritable plaisir de travailler avec eux pendant ces trois ans. Je tiens en particulier à remercier Gérard pour sa réactivité exceptionnelle, son temps de réponse aux mails avoisinant la nanoseconde.

Je remercie également le reste du jury sans qui cette soutenance n'aurait pas vu le jour. Merci donc à Peter Bülhmann et à Pierre Geurts d'avoir accepté de rapporter ma thèse. Merci pour votre lecture attentive et vos formidables remarques. Merci également à Sylvain Arlot et à Arnaud Guyader pour avoir accepté de faire partie de mon jury. Je me permets de les remercier dès à présent de ne pas m'avoir posé des questions trop complexes pendant cette soutenance et de m'avoir aidé à répondre aux questions des autres membres du jury. Je remercie également Corinne et Louise pour m'avoir permis d'organiser cette soutenance ainsi que pour leur gentillesse et leur grande efficacité tout au long de ces trois ans.

J'ai eu la chance d'effectuer ma thèse dans deux équipes différentes. Je tiens donc à remercier tous les membres du CBIO de l'institut Curie pour leur bonne humeur. Avoir une vision concrète de l'utilisation des statistiques en Machine Learning m'a beaucoup apporté. Je tiens également à adresser mes remerciements à tous les membres du LSTA qui m'ont accompagné durant cette thèse. J'ai une pensée toute particulière pour les doctorants qui ont fait vivre ce laboratoire et qui ont fait de Jussieu un endroit chaleureux et convivial.

Tout au long de ma thèse, j'ai pu rencontrer des personnes enthousiasmées par la recherche et toujours ravies de discuter et d'expliquer leurs travaux. Merci à elles, pour leur passion et pour les discussions enrichissantes qui en ont suivi.

Je remercie la fameuse Triade pour ses soirées animées, son ambiance déjantée et surtout son amitié. Je remercie également tous les TalENSiens avec qui j'ai vécu des campus inoubliables ainsi que la troupe Bloody Monday qui m'a redonné l'envie de faire du théâtre. J'espère seulement que l'ombre de ce nom ne planera pas au dessus de cette soutenance.

Je remercie ma famille, en particulier Corine et Serge, pour m'avoir soutenu pendant et bien avant le doctorat. Sans eux, je ne serais pas la personne que je suis aujourd'hui. Je remercie enfin Mathieu d'être toujours à mes côtés malgré les longues soirées passées à m'écouter disserter sur les forêts aléatoires. J'espère que notre promenade ensemble, en forêts ou non, continuera pendant encore longtemps.

Avant-propos

Les forêts comptent parmi les éléments essentiels à la pérennité de la vie sur Terre. À ce titre, il est donc primordial, sinon vital, de comprendre leur fonctionnement (mais aussi leurs doutes et leurs peines, car les forêts sont aussi des êtres sensibles). Le hasard faisant bien les choses, comprendre les mécanismes nécessaires au développement des forêts est précisément le sujet de cette thèse. Après ces préliminaires quelque peu racoleurs, pénétrons dans l'univers obscur et mystérieux des forêts aléatoires.

Les Mayas lisaient l'avenir dans les étoiles ; les scientifiques utilisent désormais les forêts aléatoires (les animaux sacrifiés, les os de poulet et les boules de cristal étant déjà pris). Même si les événements récents ont donné tort aux Mayas (le monde étant toujours entier en 2015) et raison aux scientifiques (disons plutôt qu'étant d'un naturel discret, ils n'ont jamais eu ostensiblement tort), chacun de ces groupes cherche à devenir le référent mondial en matière de prédiction. À bien regarder leurs productions, les Mayas sont largement distancés par les scientifiques : les plus récentes publications des premiers remontent à quelques milliers d'années, contre seulement quelques jours pour leurs concurrents directs. Certains savants ont donc essayé de porter le coup de grâce aux Mayas en montrant, une bonne fois pour toute, que les astres ne pouvaient pas lutter contre les forêts aléatoires. Force est de constater que, jusque là, la communauté scientifique n'a pas réussi à susciter l'adhésion des foules à la sylviculture aléatoire.

Cette thèse tente de soutenir les scientifiques dans leur tâche. Nous ne nous étendrons pas sur les raisons qui ont poussé l'auteur à prendre parti pour le groupe dominant : ces raisons lui appartiennent, chacun a droit à son jardin secret. Mais son entreprise n'a pas été vaine ! Il a prouvé au monde qu'une forêt régulièrement débroussaillée fournit des prédictions exactes ; ce à quoi les Mayas lui ont répondu que le coût d'entretien d'une forêt était astronomique. Prenant en compte la désinvolture des arboristes dans ses analyses, il a montré que, même sans ces derniers, les prédictions des forêts sont toujours outrageusement précises. Pour cela, il faut néanmoins que les arbres soient assez différents les uns des autres : aucun chercheur-paysagiste digne de ce nom n'utilise des forêts composées uniquement de chênes, d'ifs ou de boulots. La (bio)diversité, c'est la vie !

Les scientifiques pensaient ainsi avoir assis leur emprise sur le reste du monde. Cependant, contre toute attente, les Mayas argumentèrent que cette thèse ne présentait aucun intérêt car les résultats supposent que la fonction de régression est continue et prend ses valeurs dans un compact borné de $\mathbb{R}$. L'auteur dut donc retourner à ses recherches pour régler ce problème. Il y travaille encore.

Contents

Chapter 1

Introduction

Contents

1.1 Introduction

Afin d'exploiter des jeux de données dont la taille ne cesse de grandir, de nouveaux algorithmes sont nécessaires. Les forêts aléatoires, créées par L. Breiman au début des années 2000 [Breiman, 2001] font partie des algorithmes qui restent efficaces (tant d'un point de vue computationnel que prédictif) lorsqu'ils sont appliqués à des grands jeux de données. Leur construction repose sur les travaux fondateurs de Amit and Geman [1997], Ho [1998] et Dietterich [2000a] et s'appuie sur le principe de *diviser pour régner* : la forêt est composée de plusieurs arbres qui sont chacun construits avec une partie du jeu de données. La prédiction de la forêt est alors obtenue simplement en agrégeant les prédictions des arbres.

Le fait que les forêts puissent être employées pour résoudre un grand nombre de problèmes d'apprentissage a fortement contribué à leur popularité. De plus, elles ne dépendent que d'un petit nombre de paramètres faciles à calibrer. Outre leur simplicité d'utilisation (voir l'implémentation du package `R`, `randomForest`), les forêts sont également connues pour leur précision et leur capacité à traiter des jeux de données composés de peu d'observations et de nombreuses variables. Étant par ailleurs facilement parallélisables, elles font partie des méthodes permettant de traiter de grands systèmes de données réelles.

Les bons résultats des forêts dans divers domaines appliqués sont légion : dans l'environnement [voir `http://www.kaggle.com/c/dsg-hackathon` et Prasad et al., 2006, Cutler et al., 2007], en chimio-informatique [Svetnik et al., 2003], dans l'identification d'objets tridimensionnels [Shotton et al., 2011], ou encore en bioinformatique [Díaz-Uriarte and de Andrés, 2006]. Abondant dans ce sens, H. Varian, économiste en chef de Google, prône leur utilisation en économétrie dans Varian [2014]. J. Howard (Kaggle) et M. Bowles (Biomatica) vont même jusqu'à affirmer dans Howard and Bowles [2012] que *"ensembles of decision trees—often known as "random forests"—have been the most successful general-purpose algorithm in modern times"*.

Le florilège de résultats appliqués contraste avec le peu de résultats théoriques sur les forêts : bien qu'utilisées très souvent dans toute une variété de domaines, leurs propriétés mathématiques demeurent largement mal comprises. Parmi les résultats théoriques les plus célèbres figure celui de Breiman [2001] qui consiste en une borne supérieure sur le risque quadratique des forêts. Cette borne dépend à la fois de la corrélation entre les arbres et de la puissance prédictive de chaque arbre. Ce résultat a été complété par une note technique de Breiman [2004] portant sur une version simplifiée de l'algorithme original. Lin and Jeon [2006] ont ensuite établi un lien entre les forêts et les estimateurs du type plus proche voisin, qui a été étudié plus en détail par Biau and Devroye [2010]. Récemment, plusieurs articles théoriques [e.g., Biau et al., 2008, Ishwaran and Kogalur, 2010, Biau, 2012, Genuer, 2012, Zhu et al., 2012] ont porté sur des versions plus ou moins simplifiées des forêts de Breiman. Plus récemment encore, certains auteurs se sont concentrées sur des forêts très proches de l'algorithme de Breiman [2001]. Denil et al. [2013] ont prouvé le premier résultat de consistance pour les forêts en ligne. Mentch and Hooker [2014a] et Wager [2014] ont étudié la distribution limite des forêts aléatoires, lorsque le nombre d'observations et le nombre d'arbres tendent vers l'infini.

L'algorithme des forêts aléatoires est souvent considéré comme une véritable *boîte noire* qui combine de manière complexe plusieurs mécanismes difficiles à appréhender. Parmi ces mécanismes, le bagging [Breiman, 1996] et le critère de coupure des arbres CART [Classification And Regression Trees Breiman et al., 1984] ont un rôle essentiel. Le bagging (contraction de *bootstrap-aggregating*) est un schéma d'agrégation qui permet de générer des échantillons bootstrap à partir de l'échantillon initial, puis de construire un estimateur à partir de chaque échantillon, pour enfin prédire en agrégeant les estimations de chacun des arbres. Cette procédure, initialement proposée pour améliorer la robustesse des estimateurs instables, compte parmi les plus efficaces en temps de calcul, particulièrement pour des grands jeux de données où la sélection d'un bon modèle prédictif est particulièrement difficile [Bühlmann and Yu, 2002, Kleiner et al., 2012, Wager et al., 2013]. Le critère de coupure CART, quant à lui, provient du célèbre algorithme de classification et de régression CART, et est utilisé dans la construction des arbres pour sélectionner la meilleure coupure perpendiculaire aux axes. Ainsi, à chaque nœud de chaque arbre, la meilleure coupure est choisie en maximisant le critère de coupure CART basé sur l'indice de Gini (classification) ou sur l'erreur de prédiction quadratique (régression).

Bien que le bagging et le critère de coupure CART jouent un rôle central dans l'algorithme des forêts aléatoires, ils demeurent tous deux difficiles à analyser. Cela explique pourquoi la majorité des travaux théoriques ont eu pour principaux objets des versions simplifiées de l'algorithme original, supprimant notamment l'étape de rééchantillonnage des données et/ou remplaçant le critère de coupure CART et le critère d'arrêt (arrêt de l'algorithme lorsque chacune des feuilles contient un faible nombre d'observations) par des procédures plus élémentaires. La plupart des

auteurs se sont ainsi concentrés sur des modèles de forêts simplifiés qui ne prennent pas en compte l'ensemble de ces mécanismes : bien souvent, ces modèles sont construits indépendamment des données, ce qui crée un décalage entre les forêts étudiées en théorie et celles réellement utilisées.

Dans cette introduction, nous présentons le contexte théorique dans lequel s'inscrit cette thèse. Dans la Section 2, nous introduisons les notations mathématiques ainsi que l'algorithme des forêts aléatoires de Breiman [2001]. Nous présentons les premiers résultats théoriques sur les modèles simplifiés de forêts dans la Section 3. Nous regroupons, dans la Section 4, les résultats théoriques portant sur les forêts de Breiman. Enfin, les différentes contributions de la thèse sont détaillées dans la Section 5.

1.2 Construction des forêts aléatoires

1.2.1 Notations

L'objectif de cette section est de présenter l'algorithme des forêts aléatoires ainsi que les notations mathématiques utilisées tout au long de la thèse. Le cadre général est celui de la régression non paramétrique dans lequel le vecteur $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)})$ des variables explicatives est à valeurs dans $[0,1]^p$ et le but est de prédire la variable aléatoire Y à valeurs dans $\mathbb{R}$ en estimant la fonction de régression $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Pour ce faire, on suppose connu un n-échantillon $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ de variables aléatoires indépendantes et identiquement distribuées, indépendantes et de même loi que le couple $(\mathbf{X}, Y)$. Le but est alors d'utiliser le jeu de données $\mathcal{D}_n$ pour construire un estimateur $m_n : [0,1]^p \to \mathbb{R}$ de la fonction de régression m. Dans ce contexte, on dira qu'un estimateur m_n est consistant si $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \to 0$ lorsque $n \to +\infty$ (l'espérance portant sur $\mathbf{X}$ et $\mathcal{D}_n$).

Le terme forêt aléatoire désigne à la fois une collection de M arbres aléatoires et l'estimateur associé à ces M arbres. Pour le j-ème arbre de la forêt, la valeur prédite en un nouveau point $\mathbf{x}$ est notée $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$, où $\Theta_1, \ldots, \Theta_M$ sont des variables aléatoires indépendantes, distribuées selon la variable Θ, et indépendantes de $\mathcal{D}_n$. Dans la pratique, cette variable est utilisée pour rééchantillonnner le jeu de données $\mathcal{D}_n$ avant de construire chaque arbre, ainsi que pour pré-sélectionner, dans chaque cellule, un ensemble de directions admissibles pour effectuer la coupure. L'estimateur de la forêt aléatoire, qui résulte de l'agrégation des différents arbres, est construit de la manière suivante :

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n). \tag{1.1}$$

Puisque le nombre d'arbres M peut être choisi arbitrairement grand (si la puissance de calcul le permet), on peut considérer que M tend vers l'infini, et ainsi remplacer l'estimateur des forêts (1.1) par l'estimateur de la forêt infinie définie par

$$m_{\infty,n}(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}; \Theta, \mathcal{D}_n) \right], \tag{1.2}$$

où $\mathbb{E}_\Theta$ est l'espérance par rapport à Θ, conditionnellement à l'échantillon $\mathcal{D}_n$. Dans la suite, pour alléger les notations, on omettra souvent la dépendance en $\mathcal{D}_n$ (par exemple $m_{\infty,n}(\mathbf{x})$ remplacera $m_{\infty,n}(\mathbf{x}; \mathcal{D}_n)$).

1.2.2 Algorithme

Nous présentons maintenant en détail le fonctionnement de l'algorithme. Les forêts aléatoires de Breiman [2001] sont composées d'arbres dont les cellules sont des hyperrectangles dans $[0,1]^p$. À chaque étape de l'algorithme, les cellules forment une partition de $[0,1]^p$. La racine de l'arbre est $[0,1]^p$ et les nœuds terminaux (aussi appelés feuilles) forment également une partition de $[0,1]^p$. Si une feuille correspond à une région $A \subset [0,1]^p$ alors l'arbre de régression prédit pour un nouveau point $\mathbf{x} \in A$ la moyenne des Y_i associées aux observations $\mathbf{X}_i$ appartenant à A. La construction des forêts de Breiman est détaillée dans l'Algorithme 1.

L'Algorithme 1 peut sembler compliqué à première vue mais les idées sous-jacentes sont très simples. Pour mieux l'appréhender, remarquons que cet algorithme comporte trois paramètres :

1. le nombre de directions pré-sélectionnées pour couper $m_{\text{try}} \in \{1, \dots, p\}$;
2. le nombre d'observations utilisées pour construire chaque arbre $a_n \in \{1, \dots, n\}$;
3. le nombre maximal d'observations dans chaque feuille `nodesize` $\in \{1, \dots, n\}$.

L'algorithme construit M arbres aléatoires de la manière suivante. Pour chaque arbre, a_n observations sont sélectionnées avec remise parmi les n observations initiales. Chaque cellule est ensuite coupée de manière à maximiser le critère de coupure (voir plus bas) jusqu'à ce que toutes les feuilles de chaque arbre contiennent moins de `nodesize` observations.

Pour définir mathématiquement le critère de coupure, considérons une cellule A et notons $N_n(A)$ le nombre d'observations appartenant à A. Une coupure dans A est un couple (j, z), où $j \in \{1, \dots, p\}$ est la direction de la coupure et z est la position de la coupure selon la j-ème coordonnée, dans les limites de A. Soit $\mathcal{C}_A$ l'ensemble des coupures possibles dans A (i.e., qui séparent effectivement A en deux cellules non vides). En notant $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(p)})$, pour tout $(j, z) \in \mathcal{C}_A$, le critère de coupure CART s'écrit

$$\begin{aligned} L_n(j,z) =& \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\ & - \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z})^2 \mathbb{1}_{\mathbf{X}_i \in A}, \end{aligned} \tag{1.3}$$

où $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$, et $\bar{Y}_A$ (resp., $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) est la moyenne des Y_i appartenant à A (resp., A_L, A_R), avec la convention $0/0 = 0$. Pour chaque cellule A, la meilleure coupure $(j_n^\star, z_n^\star)$ est celle maximisant $L_n(j, z)$ sur l'ensemble $\mathcal{M}_{\text{try}}$ et $\mathcal{C}_A$, c'est-à-dire

$$(j_n^\star, z_n^\star) \in \underset{\substack{j \in \mathcal{M}_{\text{try}} \\ (j,z) \in \mathcal{C}_A}}{\arg\max} L_n(j, z).$$

Pour éliminer certains cas d'égalité dans l'argmax, la meilleure coupure est toujours effectuée selon une des meilleures directions $j_n^\star$ et au milieu de deux points consécutifs.

En résumé, pour chaque cellule, l'algorithme choisit uniformément `mtry` coordonnées dans $\{1, \dots, p\}$, évalue le critère (1.3) sur toutes les coupures possibles selon les `mtry` directions, et sélectionne la meilleure. Le critère (1.3) est celui utilisé dans l'algorithme CART de Breiman

Algorithm 1: Prédiction de la forêt de Breiman au point $\mathbf{x}$.

Input: Échantillon d'apprentissage $\mathcal{D}_n$, nombre d'arbres $M \in \mathbb{N}$, $\texttt{mtry} \in \{1, \dots, p\}$, $a_n \in \{1, \dots, n\}$, et $\mathbf{x} \in [0,1]^p$.
Output: Prédiction de la forêt aléatoire en $\mathbf{x}$.

for $j = 1, \dots, M$ **do**
  Tirer a_n points uniformément parmi $\mathcal{D}_n$ avec remise.
  Soit $\mathcal{P}_0 = \{[0,1]^p\}$ la partition associée à la racine de l'arbre.
  Soit $\mathcal{P}_\ell = \emptyset$, pour tout $1 \leq \ell \leq a_n$.
  Soit $n_{nodes} = 1$ et $level = 0$.
  while $n_{nodes} < a_n$ **do**
    if $\mathcal{P}_{level} = \emptyset$ **then**
      $level = level + 1$.
    else
      Soit A le premier élément de $\mathcal{P}_{level}$.
      if A *contient moins de* `nodesize` *points* **then**
        $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$.
        $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A\}$.
      else
        Tirer un sous-ensemble $\mathcal{M}_{\text{try}} \subset \{1, \dots, p\}$ de cardinal m_{try} uniformément et sans remise.
        Choisir la meilleure coupure dans la cellule A qui maximise le critère de coupure CART selon les coordonnées dans $\mathcal{M}_{\text{try}}$ *(voir les détails ci-dessous)*.
        Couper A selon la coupure précédemment choisie. Soit A_L et A_R les deux cellules filles.
        $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$.
        $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A_L\} \cup \{A_R\}$.
        $n_{nodes} = n_{nodes} + 1$.
      end
    end
  end
  Calculer la prédiction du j-ème arbre en $\mathbf{x}$, notée $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$, égale à la moyenne des Y_i appartenant à la cellule contenant $\mathbf{x}$ de la partition $\mathcal{P}_{level} \cup \mathcal{P}_{level+1}$.
end
Calculer la prédiction de la forêt aléatoire en $\mathbf{x}$, $m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n)$, donnée par la formule (1.1).

et al. [1984]. Il mesure la différence des variances dans les cellules avant et après coupure, à la différence près qu'ici le critère est évalué sur un sous ensemble des p directions. D'autre part, contrairement à l'algorithme CART original, les arbres ne sont pas élagués, et les feuilles ne contiennent qu'un petit nombre de points, qui est systématiquement inférieur à `nodesize`. De plus, chaque arbre est construit à partir d'un sous-échantillon des données de taille a_n.

La littérature est peu prolixe sur la façon dont les performances des forêts sont influencées par les paramètres de l'algorithme (M, m_{try}, a_n, `nodesize`). Cependant les paramètres par défaut semblent en général de bons choix.

1.3 Modèles de forêts aléatoires

1.3.1 Forêts non adaptatives

Comme nous l'avons souligné dans l'introduction, les forêts de Breiman reposent sur des mécanismes complexes et sont donc difficiles à analyser, notamment car leur construction dépend des données.

Par conséquent, la littérature concernant les forêts est marquée par une profonde dichotomie. Les articles appliqués décrivent des extensions parfois complexes des forêts aléatoires à divers domaines (classement, estimation de quantiles, analyse de survie...). Leurs performances sont bien souvent meilleures que celles des algorithmes classiquement utilisés dans les domaines concernés mais aucune garantie théorique ne vient corroborer ces méthodes. À l'inverse, la plupart des articles théoriques se concentrent sur des versions simplifiées de l'algorithme original pour lesquelles l'analyse théorique semble plus aisée.

Les versions simplifiées ayant pour point commun d'être construites indépendamment des données portent le nom de *purely random forests*. Parmi celles-ci, les forêts centrées ont été largement étudiées et sont construites de la manière suivante:

1. il n'y a pas d'étape de rééchantillonnage;

2. dans chaque nœud de chaque arbre, une seule coordonnée est uniformément choisie parmi $\{1, \dots, p\}$;

3. la coupure est alors effectuée au centre de la cellule selon la coordonnée précédemment choisie.

4. Les étapes $2-3$ sont répétées de façon récursive k fois ($k \in \mathbb{N}$ est un paramètre de l'algorithme), jusqu'à ce qu'un arbre complet binaire de niveau k soit obtenu (autrement dit, chaque arbre contient 2^k feuilles).

Le paramètre k est ainsi un paramètre de régularisation contrôlant le nombre de feuilles (voir Figure 1.1 pour un exemple en dimension 2). Plus k est grand, plus les arbres de la forêts seront développés, minimisant ainsi l'erreur d'approximation. À l'inverse, plus k est petit, plus les cellules sont susceptibles de contenir un grand nombre de points, minimisant ainsi l'erreur d'estimation. Les forêts uniformes sont un autre exemple de *purely random forests*. Elles sont construites de la même manière que les forêts centrées hormis le fait qu'une fois la direction de coupure sélectionnée, la position de coupure est choisie uniformément selon cette direction (dans les limites de la cellule). Bien que leur construction diffère de celle des forêts centrées, leur analyse n'en demeure pas moins largement similaire.

Les forêts centrées ont été formalisées par Biau [2012]. Dans cet article, l'auteur prouve que ces estimateurs sont consistants si $k \to +\infty$ et $n/2^k \to +\infty$. L'hypothèse $k \to +\infty$ permet

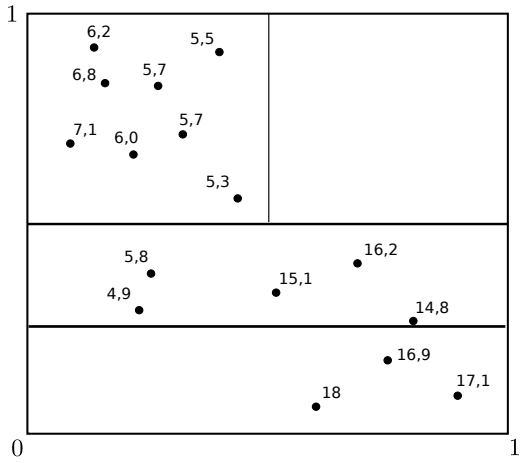

Figure 1.1: Un arbre centré de niveau 2.

de contrôler l'erreur d'approximation de la forêt en assurant que les arbres sont assez développés. D'autre part, si $\mathbf{X}$ est uniformément distribué sur $[0,1]^p$, alors il y a $n/2^k$ observations en moyenne dans chaque feuille. L'hypothèse $n/2^k \to +\infty$ permet de contrôler l'erreur d'estimation de la forêt en assurant que les feuilles contiennent un grand nombre d'observations. Par conséquent, les forêts centrées ne constituent pas un modèle adéquat des forêts de Breiman car les feuilles de ces dernières ne contiennent qu'un petit nombre d'observations. De plus, la consistance des forêts centrées résulte de la consistance des arbres qui les composent. En résumé, ce modèle ne permet pas de mettre en lumière les avantages des forêts par rapport aux arbres de régression : appréhender les mécanismes en œuvre dans les forêts de Breiman nécessite d'étudier des modèles plus complexes.

Étant plus faciles à analyser que les forêts de Breiman, les *purely random forests* sont à cet égard des modèles intéressants à étudier. En effet, la vitesse de convergence des forêts centrées a été déterminée par Breiman [2004] et Biau [2012]. Dans leur approche, on se donne un ensemble $\mathcal{S} \subset \{1,\dots,p\}$. La fonction de régression $m(\mathbf{X})$, qui est initialement une fonction de $\mathbf{X} = (\mathbf{X}^{(1)},\dots,\mathbf{X}^{(p)})$ est supposée ne dépendre que des $|\mathcal{S}|$ (où $|\mathcal{S}|$ désigne le cardinal de $\mathcal{S}$) coordonnées de $\mathcal{S}$. Les variables restantes, c'est-à-dire celles appartenant à l'ensemble $\{1,\dots,p\}\setminus \mathcal{S}$, n'ont aucune influence sur la fonction de régression qui se réécrit alors

$$m(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}_{\mathcal{S}}],$$

où $\mathbf{X}_{\mathcal{S}} = (\mathbf{X}^{(j)} : j \in \mathcal{S})$. Breiman [2004] et Biau [2012] ont prouvé que si les arbres sont construits en utilisant uniquement les variables importantes (i.e., celles appartenant à $\mathcal{S}$) et si m est lipschitzienne alors

$$\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = \mathrm{O}\left(n^{\frac{-0.75}{|\mathcal{S}|\log 2 + 0.75}}\right). \tag{1.4}$$

Ce résultat montre que la vitesse de convergence des forêts centrées ne dépend que du nombre de variables importantes $\mathcal{S}$ et non de la dimension ambiante p. De plus, la vitesse de convergence des

forêts centrées (1.4) est plus rapide que la vitesse minimax $n^{-2/(p+2)}$ dès lors que $|\mathcal{S}| \leq \lfloor 0.54p \rfloor$. Ce résultat, perturbant de prime abord, n'est pas absurde. En effet, la dimension intrinsèque du problème est $|\mathcal{S}|$ et non p. La véritable vitesse minimax du problème est donc $n^{-2/(|\mathcal{S}|+2)}$. Les forêts sont donc capables de s'adapter à la dimension du problème (car leur vitesse est meilleure que la vitesse $n^{-2/(p+2)}$) mais ne dépassent pas la vitesse minimax correspondant au véritable problème de régression. Ce phénomène peut s'avérer particulièrement utile lorsque le nombre de variables importantes est beaucoup plus petit que le nombre total de variables (cadre de la parcimonie en grande dimension). Ce résultat peut aussi permettre d'expliquer pourquoi les forêts aléatoires ne sur-ajustent pas, même lorsque le nombre de variables est grand.

Un autre modèle de *purely random forest* est la *Purely Uniform Random Forest* (PURF) étudiée par Genuer [2012]. Pour $p = 1$, une PURF est obtenue en sélectionnant k variables aléatoires uniformes sur $[0, 1]$ et en divisant $[0, 1]$ en $k + 1$ sous-intervalles obtenus grâce à ses variables. Bien que cette construction ne soit pas récursive, elle est équivalente à la construction d'un arbre de décision dans lequel, à chaque étape, on choisit quelle cellule couper avec une probabilité égale à sa taille. Genuer [2012] a montré que les PURF sont consistantes et que si la fonction de régression est lipschitzienne, alors

$$\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 = \mathrm{O}\left(n^{-2/3}\right),$$

ce qui correspond à la vitesse minimax sur la classe des fonction lipschitziennes [Stone, 1980, 1982]. Ce résultat est rassurant : les forêts sont capables dans ce contexte particulier d'être au moins aussi performantes asymptotiquement que divers estimateurs plus classiques (estimateurs à noyau par exemple).

Il est communément admis que l'action d'agréger des arbres permet de réduire l'erreur d'estimation des arbres individuels tout en conservant une erreur d'approximation du même ordre de grandeur. Biau [2012] souligne que l'erreur d'estimation des forêts centrées tend vers zero (à une faible vitesse $1/\log n$) lorsque chaque arbre est complètement développé (i.e., $k \approx \log n$). Cependant l'erreur d'estimation d'un arbre complètement développé ne tend pas vers 0. Le processus d'agrégation d'arbres permet ainsi de réduire drastiquement l'erreur d'estimation. Malheureusement le choix $k \approx \log n$ est trop grand pour assurer la consistance de la forêt centrée (l'erreur d'approximation étant alors constante). De la même manière, Genuer [2012] souligne que l'erreur d'estimation des forêts PURF est multipliée par un facteur 0.75 par rapport à l'erreur d'estimation des arbres individuels. La tentative la plus récente pour étudier à la fois l'erreur d'estimation et d'approximation d'une forêt provient d'Arlot and Genuer [2014], qui mettent en évidence plusieurs modèles de forêts dont l'erreur d'approximation est plus faible que l'erreur d'approximation des arbres individuels qui les constituent.

1.3.2 Plus proches voisins

Les forêts aléatoires peuvent également être rapprochées de certains algorithmes à moyennes locales (méthode des plus proches voisins, estimateurs à noyaux). Pour établir plus précisément ce lien, nous avons besoin de la notion de *potentiel plus proche voisin*. En géométrie aléatoire, une observation $\mathbf{X}_i$ est appelée un *potentiel plus proche voisin* (PPPV) d'un point $\mathbf{x}$ (parmi $\mathbf{X}_1, \ldots, \mathbf{X}_n$) si l'hyperrectangle défini par $\mathbf{x}$ et $\mathbf{X}_i$ ne contient pas d'autres points (Barndorff-Nielsen and Sobel, 1966, Bai et al., 2005; voir aussi Devroye et al., 1996, Chapitre 11, Problème

6). Comme le montre la Figure 1.2, le nombre de PPPV de $\mathbf{x}$ est en général plus grand que 1 et dépend à la fois du nombre d'observations et de leur répartition. Il s'avère que le concept de

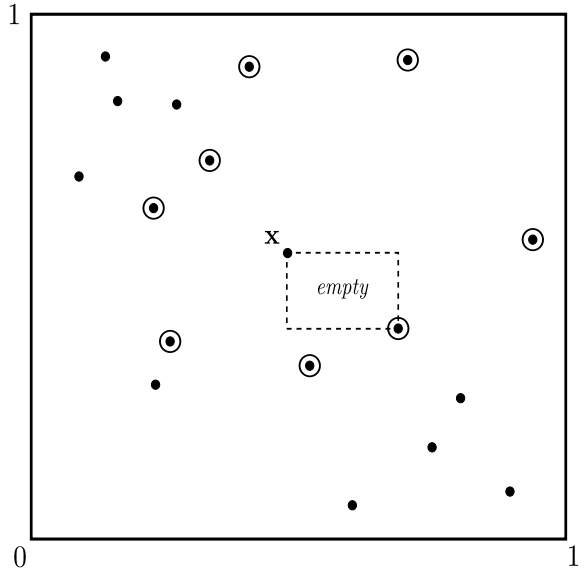

Figure 1.2: Potentiels plus proches voisins (PPPV) de $\mathbf{x}$ en dimension $p = 2$.

PPPV est intimement lié aux forêts aléatoires. En effet, si les feuilles de chaque arbre contiennent exactement une observation, alors quelle que soit la stratégie de coupure utilisée, l'estimateur des forêts aléatoires est une moyenne pondérée des Y_i associés aux $\mathbf{X}_i$ étant des PPPV de $\mathbf{x}$. Mathématiquement, on a alors

$$m_{\infty,n}(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) Y_i, \tag{1.5}$$

où les poids $(W_{n1}, \ldots, W_{nn})$ sont des fonctions positives de l'échantillon $\mathcal{D}_n$ et satisfont la contrainte $W_{ni}(\mathbf{x}) = 0$ si $\mathbf{X}_i$ n'est pas un PPPV de $\mathbf{x}$. Remarquant cette intéressante connexion entre les forêts et les PPPV, Lin and Jeon [2006] ont établi que, si $\mathbf{X}$ est uniformément distribué sur $[0,1]^p$ et si les arbres sont construits indépendamment des valeurs $Y_1, \ldots, Y_n$, alors

$$\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = \mathrm{O}\left(\frac{1}{n_{\max}(\log n)^{p-1}}\right),$$

où $n_{\max}$ est le nombre maximal de points dans les feuilles (Biau and Devroye, 2010 ont étendu cette inégalité au cas où $\mathbf{X}$ admet une densité quelconque sur $[0,1]^p$). Malheureusement, les valeurs exacts des poids $W_{n1}, \ldots, W_{nn}$ de la forêt de Breiman sont inconnus, et il semble donc difficile de réécrire l'estimateur (1.5) sous une forme plus explicite. Néanmoins, les poids W_{ni} sont reliés à la probabilité de connexion entre deux points $\mathbf{x}$ et $\mathbf{z}$ définie par

$$K_n(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta\left[\mathbf{z} \in A_n(\mathbf{x}, \Theta)\right],$$

où $A_n(\mathbf{x}, \Theta)$ est la cellule contenant $\mathbf{x}$ de l'arbre construit avec le paramètre Θ, et $\mathbb{P}_\Theta$ est la probabilité par rapport à Θ, conditionnellement à l'échantillon $\mathcal{D}_n$. Ce lien, étudié plus en détail dans le **Chapitre 4**, nous permettra de donner une forme explicite approchée de certains estimateurs des forêts aléatoires.

1.4 Résultats théoriques sur les forêts de Breiman

Rééchantillonnage L'étape de ré-échantillonnage dans l'algorithme de Breiman [2001] s'effectue en choisissant n observations parmi n avec remise pour construire chaque arbre. Cette procédure, initialement proposée par Efron [1982] [voir également Politis et al., 1999], est appelée bootstrap dans la littérature statistique. L'idée de générer de nombreux échantillons bootstrap et d'agréger les estimateurs produits à partir de ces échantillon est appelée bagging. Cette méthode fut suggérée par Breiman [1996] comme un moyen simple d'améliorer la précision d'estimateurs instables ou peu performants. Bien qu'un des grands avantages du bootstrap soit sa simplicité d'utilisation, la théorie le concernant s'avère plus compliquée. En effet, les observations bootstrappées ont une distribution différente de celle l'échantillon initial.

Le rôle du bootstrap dans les forêts aléatoires demeure mal compris et, à ce jour, la plupart des analyses remplacent le bootstrap par du sous-échantillonnage en supposant que chaque arbre est construit avec $a_n < n$ observations, sélectionnées aléatoirement parmi l'échantillon initial [Wager, 2014, Mentch and Hooker, 2014a]. La plupart du temps, le taux de sous-échantillonnage a_n/n est supposé tendre vers zero à une vitesse spécifique, une hypothèse qui exclut de fait le cas du bootstrap où $a_n = n$.

Positions des coupures Le procédé de coupure n'est pas facile à appréhender dans la mesure où il utilise à la fois les positions $\mathbf{X}_i$ et les valeurs Y_i pour sélectionner la meilleure coupure. A partir des idées développées par Bühlmann and Yu [2002], Banerjee and McKeague [2007] ont établi la loi limite pour la position de la coupure optimale dans le cadre d'un modèle de régression de la forme $Y = m(\mathbf{X}) + \varepsilon$, où $\mathbf{X}$ est une variable aléatoire réelle et ε est un bruit Gaussien indépendant. Afin d'énoncer leur résultat, supposons pour l'instant que la distribution de $(\mathbf{X}, Y)$ est connue et notons $d^\star$ la coupure optimale qui maximise l'équivalent théorique $L^\star$ du critère CART empirique (1.3) défini, dans toute cellule A par

$$\begin{aligned} L^\star(j,z) = \mathbb{V}\big[Y|\mathbf{X} \in A\big] \\ - \mathbb{P}\big[\mathbf{X} \in A, \mathbf{X}^{(j)} \leq z\big]\mathbb{V}\big[Y|\mathbf{X} \in A, \mathbf{X}^{(j)} \leq z\big] \\ - \mathbb{P}\big[\mathbf{X} \in A, \mathbf{X}^{(j)} > z\big]\mathbb{V}\big[Y|\mathbf{X} \in A, \mathbf{X}^{(j)} > z\big]. \end{aligned}$$

Dans ce modèle, l'estimateur vaut respectivement, sur chacune des deux cellules engendrées,

$$\beta_\ell^\star = \mathbb{E}[Y|X \leq d^\star] \quad \text{and} \quad \beta_r^\star = \mathbb{E}[Y|X > d^\star].$$

Bien entendu, quand la distribution du couple $(\mathbf{X}, Y)$ est inconnue, il en va de même pour les quantités $\beta_\ell^\star, \beta_r^\star, d^\star$ qui sont estimées par leurs contreparties empiriques

$$(\hat{\beta}_\ell, \hat{\beta}_r, \hat{d}_n) \in \underset{\beta_\ell, \beta_r, d}{\arg\min} \sum_{i=1}^{n} \big[Y_i - \beta_\ell \mathbb{1}_{X_i \leq d} - \beta_r \mathbb{1}_{X_i > d}\big]^2.$$

Si le modèle de régression possède certaines régularités (entre autres, $\mathbf{X}$ a une densité C^1 et m est C^1), Banerjee and McKeague [2007] ont prouvé que

$$n^{1/3}(\hat{\beta}_l - \beta_l^\star, \hat{\beta}_r - \beta_r^\star, \hat{d}_n - d^\star) \xrightarrow{\mathcal{D}} (c_1, c_2, 1) \arg\max_t Q(t), \tag{1.6}$$

où $\mathcal{D}$ correspond à la convergence en loi, $Q(t) = aW(t) - bt^2$, et W est un mouvement Brownien standard réel (les constantes a et b sont positives et dépendent des paramètres du modèle et des quantités inconnues $\beta_\ell^\star, \beta_r^\star$ et $d^\star$). La loi limite (1.6) permet de construire des intervalles de confiance asymptotique pour les quantités $\beta_\ell^\star$, $\beta_r^\star$ et $d^\star$, ce qui peut s'avérer particulièrement intéressant dans certains problèmes où la position des points de rupture admet une interprétation simple [voir par exemple Banerjee and McKeague, 2007].

Une autre analyse de la position des coupures a été réalisée par Ishwaran [2013]. L'auteur s'est intéressé à l'ECP (*End-Cut Preference*) du critère de coupure CART, c'est-à-dire au fait que les coupures portant sur des variables non informatives se concentrent avec grande probabilité autour des bords des cellules [ce phénomène avait déjà été observé dans Breiman et al., 1984]. Ishwaran [2013] a souligné l'aspect positif de cette propriété qui était jusque là perçue comme un défaut de l'algorithme CART. Pour bien la comprendre, considérons une cellule A contenant n points et supposons que cette cellule est coupée en son centre selon une variable non informative. Alors les deux cellules créées contiennent environ la moitié des données (si $\mathbf{X}$ est uniformément réparti sur A) mais, puisque la variable est non informative, la coupure ne permet pas de réduire l'erreur d'approximation dans chacune des cellules. Dans ce cas, l'erreur d'estimation a donc augmenté car la taille de l'échantillon a été divisée par deux dans chacune des deux cellules engendrées. On obtient donc une configuration où la coupure a augmenté l'erreur d'estimation en laissant inchangée l'erreur d'approximation. L'ECP permet d'éviter cette configuration en coupant préférentiellement les variables non informatives près du bord des cellules.

Forêts de Breiman Tout compte fait, peu de résultats théoriques portent sur les forêts de Breiman [2001]. Wager [2014] et Mentch and Hooker [2014a], ont chacun établi dans des contextes légèrement différents la normalité asymptotique des estimateurs des forêts aléatoires. Wager [2014] a montré que la variance des forêts infinies pouvait être estimée de manière consistante grâce au Jackknife, ce qui permet d'évaluer la qualité des prédictions des forêts aléatoires. Mentch and Hooker [2014a] ont démontré le même type de résultat pour des forêts avec un nombre fini d'arbres, en distinguant plusieurs régimes asymptotiques.

1.5 Contributions

Nous avons choisi de structurer notre travail en cinq chapitres qui peuvent être lus indépendamment les uns des autres. Le **Chapitre 2** est une revue de la littérature sur les aspects théoriques des forêts aléatoires, coécrit avec Gérard Biau et soumis au journal *TEST*. Le **Chapitre 3** étudie certaines propriétés asymptotiques des forêts aléatoires et a été accepté pour publication dans la revue *Journal of Multivariate Analysis*. Le **Chapitre 4** porte sur le lien entre les forêts et les méthodes à noyau et a été soumis à la revue *IEEE Transactions on Information Theory*. Le **Chapitre 5**, coécrit avec Gérard Biau et Jean-Philippe Vert, a pour objet la consistance des

forêts de Breiman et a été accepté pour publication dans le journal *The Annals of Statistics*. Enfin le **Chapitre 6**, relativement indépendant par son contenu des autres chapitres, présente une étude de cas réalisée avec l'équipe du CBIO (Centre for computational BIOlogy) de l'Institut Curie.

Chapitre 2

Nous proposons dans le **Chapitre 2** une revue de la littérature sur la théorie des forêts aléatoires, venant compléter la présente introduction. Elle inclut notamment les résultats sur la sélection de variables et sur certaines extensions algorithmiques des forêts aléatoires.

Chapitre 3

Avant d'établir certaines propriétés des forêts aléatoires (que ce soient celles de Breiman ou d'autres modèles simplifiés), il nous paraissait nécessaire d'étudier plus en détail le lien entre les forêts infinies (analysées en théorie, voir équation (1.2)) et les forêts finies (utilisées en pratique, voir équation (1.1)). C'est l'objet de la première partie de ce chapitre. Nous montrons une loi des grands nombres uniforme en le point d'estimation $\mathbf{x}$ reliant l'estimateur des forêts finies et celui des forêts infinies et valable pour une grande classe de modèles de forêts (**Théorème 3.1**). Une analyse plus poussée nous permet ensuite de déduire un théorème central limite uniforme en $\mathbf{x}$ pour les estimateurs des forêts finies (**Théorème 3.2**). Enfin, nous montrons dans le **Théorème 3.3** (reproduit ci-dessous) que, sous certaines hypothèses sur le modèle de régression, l'erreur $\mathbb{L}^2$ des forêts finies est proche de celle des forêts infinies, si le nombre d'arbres est bien choisi.

Théorème 3.3. *Supposons que*

$$Y = m(\mathbf{X}) + \varepsilon,$$

où ε est une bruit Gaussien centré avec une variance $\sigma^2 < +\infty$, indépendante de $\mathbf{X}$, et $\|m\|_\infty = \sup_{\mathbf{x}\in[0,1]^p} |m(\mathbf{x})| < \infty$. Alors, pour tout $M, n \in \mathbb{N}^\star$,

$$0 \leq R(m_{M,n}) - R(m_{\infty,n}) \leq \frac{8}{M} \times (\|m\|_\infty^2 + \sigma^2(1 + 4\log n)),$$

où, pour tout estimateur m_n, $R(m_n) = \mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2$.

En particulier, ce résultat permet d'étendre la consistance des forêts infinies aux forêts finies sous les hypothèses précédentes sur le modèle de régression et en choisissant le nombre d'arbres M de sorte que $M/\log n \to +\infty$. Ces résultats nous permettent de nous concentrer dans le reste de la thèse sur les propriétés des forêts infinies.

Dans la deuxième partie du **Chapitre 3**, nous prouvons un théorème général sur la consistance des forêts aléatoires dont la construction est indépendante des données (**Théorème 3.4**). Dans ce théorème, comme dans les exemples mentionnés à la Section 3.1 de cette introduction, la consistance de la forêt résulte de la consistance des arbres individuels qui la composent. Afin de mettre en exergue certaines propriétés propres aux forêts aléatoires, nous considérons

un modèle plus proche des forêts de Breiman : les forêts quantiles dont les nœuds terminaux contiennent exactement une observation (voir Algorithme 1, **Chapitre 3**). Dans ce contexte, nous montrons dans le **Théorème 3.5** que le sous-échantillonnage est crucial pour assurer la consistance de la forêt. En effet, les arbres de la forêt quantile sont inconsistants (car leurs nœuds terminaux ne contiennent qu'un seul point) mais la forêt quantile est consistante grâce au sous-échantillonnage.

Théorème 3.5. *Supposons que*

$$Y = m(\mathbf{X}) + \varepsilon,$$

où ε est un bruit centré vérifiant $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \leq \sigma^2$, avec $\sigma^2 < \infty$ une constante. Supposons de plus que $\mathbf{X}$ admet une densité sur $[0,1]^p$ et que m est continue. Alors, si $a_n \to +\infty$ et $a_n/n \to 0$, la forêt infinie q quantile est consistante.

Chapitre 4

Afin de trouver une forme explicite facilement interprétable de l'estimateur des forêts aléatoires, nous présentons dans le **Chapitre 4** un lien entre les forêts infinies et les méthodes à noyau. En modifiant légèrement la manière dont les arbres sont agrégés par la forêt, l'estimateur des forêts aléatoires peut se réécrire comme un estimateur à noyau de la forme

$$\tilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_n(\mathbf{X}_i, \mathbf{x})}{\sum_{j=1}^{n} K_n(\mathbf{X}_j, \mathbf{x})}, \tag{1.7}$$

où $K_n(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta [\mathbf{z} \in A_n(\mathbf{x}, \Theta)]$ est la probabilité de connexion de $\mathbf{x}$ et de $\mathbf{z}$ dans la forêt, c'est-à-dire la probabilité que $\mathbf{x}$ et $\mathbf{z}$ soient dans la même cellule d'un arbre aléatoire de la forêt (**Proposition 4.2**). Le noyau K_n correspond donc à une mesure de proximité particulière, intrinsèque à la forêt considérée. Les estimateurs de la forme (1.7) ont des performances similaires (que ce soit en précision ou en temps de calcul) à celles des forêts originales, tout en étant plus facilement interprétables. D'autre part, pour certains modèles de forêts, la probabilité de connexion K_n peut être explicitée (**Propositions 4.5 et 4.6**), ce qui nous permet d'obtenir des bornes sur les vitesses de convergences des estimateurs $\tilde{m}_{\infty,n}$ (**Théorèmes 4.1 et 4.2**).

Chapitre 5

Le **Chapitre 5** est consacré aux principaux résultats de cette thèse sur les forêts aléatoires de Breiman. Nous énonçons deux théorèmes sur la consistance des forêts de Breiman dans le cadre d'un modèle de régression additif. Le premier porte sur des forêts aléatoires élaguées (i.e., non complètement développées) et repose sur le fait que les arbres individuels sont consistants, quel que soit le taux de sous-échantillonnage. Plus précisément, si on suppose que chaque arbre de la forêt de Breiman est construit à partir de a_n observations (sous-échantillonnage) et contient au maximum t_n feuilles (procédure d'élagage) alors le **Théorème 5.1** est vérifié.

Théorème 5.1. *Sous certaines hypothèses sur le modèle de régression, si $a_n \to +\infty$, $t_n \to +\infty$ et si $t_n(\log a_n)^9/a_n \to 0$, les forêts de Breiman sont consistantes.*

La preuve du **Théorème 5.1** montre également que chaque arbre de la forêt est consistant. Autrement dit, les arbres CART, élagués de cette façon, sont consistants. Le **Théorème 5.2** quant à lui concerne les forêts de Breiman complètement développées et suppose un taux de sous-échantillonnage bien choisi (comme pour les forêts quantiles).

Théorème 5.2. *Sous les mêmes hypothèses que précédemment, si* **(H5.2)** *est vérifiée et si* $a_n \to +\infty$ *et* $a_n \log n/n \to 0$*, alors les forêts de Breiman complètement développées (i.e.,* $t_n = a_n$*) sont consistantes.*

Malheureusement, le **Théorème 5.2** s'appuie sur une conjecture **(H5.2)** portant sur la faible dépendance des arbres individuels en l'échantillon d'apprentissage qui semble difficile à vérifier. Ces résultats de consistance sont les premiers de ce type pour l'algorithme original de Breiman [2001].

Chapitre 6

Nous avons participé avec les membres du CBIO de l'institut Curie (Elsa Bernard, Yunlong Jiao, Veronique Stoven, Thomas Walter et Jean-Philippe Vert) à l'un des Dreamchallenge (`http://dreamchallenges.org/`) de 2013 organisés par le consortium SAGE. Lors des Dreamchallenge, un problème d'apprentissage est posé et des données sont mises à disposition des participants, qui tentent alors de proposer des solutions innovantes pour résoudre le problème initial. Le **Chapitre 6** présente les résultats que nous avons obtenus à un Dreamchallenge visant à prédire la toxicité de certains composés chimiques en fonction du profil génétique de chaque individu. Différents types de variables descriptives pour chaque patient (sexe, ethnie, séquence ARN, SNP) étaient mis à disposition ainsi que plusieurs types de variables descriptives pour chacun des 156 composés chimiques dont la toxicité était évaluée au moyen de l'EC10 (mesure de la concentration du composé pour laquelle le niveau d'ATP de la cellule est réduit de 10%). Pour résoudre ce problème de régression multi-tâches, nous avons utilisé des méthodes à noyau qui permettent de prendre en compte les différents types de données. La proximité entre deux profils génétiques a été mesurée grâce à un noyau intégrant les différents types d'informations et la proximité entre les différents composés a été évaluée grâce à un noyau empirique. Cette méthode s'est classée deuxième parmi les cent méthodes proposées par l'ensemble des candidats.

Chapter 2

A Random Forest Guided Tour

Abstract The random forest algorithm, proposed by L. Breiman in 2001, has been extremely successful as a general purpose classification and regression method. The approach, which combines several randomized decision trees and aggregates their predictions by averaging, has shown excellent performance in settings where the number of variables is much larger than the number of observations. Moreover, it is versatile enough to be applied to large-scale problems, is easily adapted to various ad-hoc learning tasks, and returns measures of variable importance. The present article reviews the most recent theoretical and methodological developments for random forests. Emphasis is placed on the mathematical forces driving the algorithm, with special attention given to the selection of parameters, the resampling mechanism, and variable importance measures. This review is intended to provide non-experts easy access to the main ideas.

Contents

2.1 Introduction

To take advantage of the sheer size of modern data sets, we now need learning algorithms that scale with the volume of information, while maintaining sufficient statistical efficiency. Random forests, devised by L. Breiman in the early 2000s [Breiman, 2001], are part of the list of the most successful methods currently available to handle data in these cases. This supervised learning procedure, influenced by the early work of Amit and Geman [1997], Ho [1998], and Dietterich [2000b], operates according to the simple but effective "divide and conquer" principle: sample small fractions of the data, grow a randomized tree predictor on each small piece, then paste (aggregate) these predictors together.

What has greatly contributed to the popularity of forests is the fact that they can be applied to a wide range of prediction problems and have few parameters to tune. Aside from being simple to use, the method is generally recognized for its accuracy and its ability to deal with small sample sizes and high-dimensional feature spaces. At the same time, it is easily parallelizable and has therefore the potential to deal with large real-life systems. The corresponding `R` package `randomForest` can be freely downloaded on the CRAN website (http://www.r-project.org), while a MapReduce [Jeffrey and Sanja, 2008] open source implementation called *Partial Decision Forests* is available on the Apache Mahout website at https://mahout.apache.org. This allows the building of forests using large data sets as long as each partition can be loaded into memory.

The random forest methodology has been successfully involved in various practical problems, including a data science hackathon on air quality prediction (http://www.kaggle.com/c/dsg-hackathon), chemoinformatics [Svetnik et al., 2003], ecology [Prasad et al., 2006, Cutler et al., 2007], 3D object recognition [Shotton et al., 2011] and bioinformatics [Díaz-Uriarte and de Andrés, 2006], just to name a few. J. Howard (Kaggle) and M. Bowles (Biomatica) claim in Howard and Bowles [2012] that *ensembles of decision trees—often known as "random forests"—have been the most successful general-purpose algorithm in modern times*, while H. Varian, Chief Economist at Google, advocates in Varian [2014] the use of random forests in econometrics.

On the theoretical side, the story of random forests is less conclusive and, despite their extensive use, little is known about the mathematical properties of the method. The most celebrated theoretical result is that of Breiman [2001], which offers an upper bound on the generalization error of forests in terms of correlation and strength of the individual trees. This was followed by a technical note [Breiman, 2004], which focuses on a stylized version of the original algorithm [see also Breiman, 2000a,b]. A critical step was subsequently taken by Lin and Jeon [2006], who highlighted an interesting connection between random forests and a particular class of nearest neighbor predictors, further developed by Biau and Devroye [2010]. In recent years, various theoretical studies have been performed [e.g., Meinshausen, 2006, Biau et al., 2008, Ishwaran and Kogalur, 2010, Biau, 2012, Genuer, 2012, Zhu et al., 2012], analyzing more elaborate models and moving ever closer to the practical situation. Recent attempts towards narrowing the gap between theory and practice include that of Denil et al. [2013], who prove the first consistency result for online random forests, and Mentch and Hooker [2014a] and Wager [2014], who study the asymptotic distribution of forests.

The difficulty in properly analyzing random forests can be explained by the black-box flavor of the method, which is indeed a subtle combination of different components. Among the forests' essential ingredients, both bagging [Breiman, 1996] and the Classification And Regression

Trees (CART)-split criterion [Breiman et al., 1984] play critical roles. Bagging (a contraction of bootstrap-aggregating) is a general aggregation scheme, which generates bootstrap samples from the original data set, constructs a predictor from each sample, and decides by averaging. It is one of the most effective computationally intensive procedures to improve on unstable estimates, especially for large, high-dimensional data sets, where finding a good model in one step is impossible because of the complexity and scale of the problem [Bühlmann and Yu, 2002, Kleiner et al., 2012, Wager et al., 2013]. As for the CART-split criterion, it originates from the influential CART algorithm of Breiman et al. [1984], and is used in the construction of the individual trees to choose the best cuts perpendicular to the axes. At each node of each tree, the best cut is selected by optimizing the CART-split criterion, based on the so-called *Gini impurity* (for classification) or the prediction squared error (for regression).

However, while bagging and the CART-splitting scheme play key roles in the random forest mechanism, both are difficult to analyze with rigorous mathematics, thereby explaining why theoretical studies have so far considered simplified versions of the original procedure. This is often done by simply ignoring the bagging step and/or replacing the CART-split selection by a more elementary cut protocol. As well as this, in Breiman's [2001] forests, each leaf (that is, a terminal node) of individual trees contains a fixed pre-specified number of observations (this parameter is usually chosen between 1 and 5). Disregarding the subtle combination of all these components, most authors have focused on stylized, data-independent procedures, thus creating a gap between theory and practice.

The goal of this survey is to embark the reader on a guided tour of random forests. We focus on the theory behind the algorithm, trying to give an overview of major theoretical approaches while discussing their inherent pros and cons. For a more methodological review covering applied aspects of random forests, we refer to the surveys by Criminisi et al. [2011] and Boulesteix et al. [2012]. We start by gently introducing the mathematical context in Section 2 and describe in full detail Breiman's [2001] original algorithm. Section 3 focuses on the theory for a simplified forest model called *purely random forests*, and emphasizes the connections between forests, nearest neighbor estimates and kernel methods. Section 4 provides some elements of theory about resampling mechanisms, the splitting criterion and the mathematical forces at work in Breiman's approach. Section 5 is devoted to the theoretical aspects of associated variable selection procedures. Lastly, Section 6 discusses various extensions to random forests including online learning, survival analysis and clustering problems.

2.2 The random forest estimate

2.2.1 Basic principles

As mentioned above, the random forest mechanism is versatile enough to deal with both supervised classification and regression tasks. However, to keep things simple, we focus in this introduction on regression analysis, and only briefly survey the classification case.

Our goal in this section is to provide a concise but mathematically precise presentation of the algorithm for building a random forest. The general framework is nonparametric regression estimation, in which an input random vector $\mathbf{X} \in [0,1]^p$ is observed, and the goal is to predict the square integrable random response $Y \in \mathbb{R}$ by estimating the regression function

$m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. With this aim in mind, we assume we are given a training sample $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of independent random variables distributed the same as the independent prototype pair $(\mathbf{X}, Y)$. The goal is to use the data set $\mathcal{D}_n$ to construct an estimate $m_n : [0,1]^p \to \mathbb{R}$ of the function m. In this respect, we say that the regression function estimate m_n is (mean squared error) consistent if $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \to 0$ as $n \to \infty$ (the expectation is evaluated over $\mathbf{X}$ and the sample $\mathcal{D}_n$).

A random forest is a predictor consisting of a collection of M randomized regression trees. For the j-th tree in the family, the predicted value at the query point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed the same as a generic random variable Θ and independent of $\mathcal{D}_n$. In practice, the variable Θ is used to resample the training set prior to the growing of individual trees and to select the successive directions for splitting—more precise definitions will be given later. At this stage, we note that the trees are combined to form the (finite) forest estimate

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n). \tag{2.1}$$

In the `R` package `randomForest`, the default value of M (the number of trees in the forest) is `ntree` $= 500$. Since M may be chosen arbitrarily large (limited only by available computing resources), it makes sense, from a modeling point of view, to let M tends to infinity, and consider instead of (2.1) the (infinite) forest estimate

$$m_{\infty,n}(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}; \Theta, \mathcal{D}_n) \right].$$

In this definition, $\mathbb{E}_\Theta$ denotes the expectation with respect to the random parameter Θ, conditional on $\mathcal{D}_n$. In fact, the operation "$M \to \infty$" is justified by the law of large numbers, which asserts that almost surely, conditional on $\mathcal{D}_n$,

$$\lim_{M \to \infty} m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = m_{\infty,n}(\mathbf{x}; \mathcal{D}_n)$$

(see for instance Breiman, 2001, and Scornet, 2014, for more information on this limit calculation). In the following, to lighten notation we will simply write $m_{\infty,n}(\mathbf{x})$ instead of $m_{\infty,n}(\mathbf{x}; \mathcal{D}_n)$.

Classification. In the (binary) supervised classification problem [Devroye et al., 1996], the random response Y takes values in $\{0, 1\}$ and, given $\mathbf{X}$, one has to guess the value of Y. A classifier or classification rule m_n is a Borel measurable function of $\mathbf{x}$ and $\mathcal{D}_n$ that attempts to estimate the label Y from $\mathbf{x}$ and $\mathcal{D}_n$. In this framework, one says that the classifier m_n is consistent if its conditional probability of error

$$L(m_n) = \mathbb{P}[m_n(\mathbf{X}) \neq Y | \mathcal{D}_n]$$

satisfies

$$\lim_{n \to \infty} \mathbb{E} L(m_n) = L^\star,$$

where $L^\star$ is the error of the optimal—but unknown—Bayes classifier:

$$m^\star(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}[Y = 1 | \mathbf{X} = \mathbf{x}] > \mathbb{P}[Y = 0 | \mathbf{X} = \mathbf{x}] \\ 0 & \text{otherwise.} \end{cases}$$

In the classification situation, the random forest classifier is obtained via a majority vote among the classification trees, that is,

$$m_{M,n}(\mathbf{x};\Theta_1,\ldots,\Theta_M,\mathcal{D}_n) = \begin{cases} 1 & \text{if } \frac{1}{M}\sum_{j=1}^{M} m_n(\mathbf{x};\Theta_j,\mathcal{D}_n) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

2.2.2 Algorithm

We now provide some insight on how the individual trees are constructed and how randomness kicks in. In Breiman's [2001] original forests, each node of a single tree is associated with a hyperrectangular cell. At each step of the tree construction, the collection of cells forms a partition of $[0,1]^p$. The root of the tree is $[0,1]^p$ itself, and the terminal nodes (or leaves), taken together, form a partition of $[0,1]^p$. If a leaf represents region A, then the regression tree outputs on A the average of all Y_i for which the corresponding $\mathbf{X}_i$ falls in A. Algorithm 2 describes in full detail how to compute a forest's prediction.

Algorithm 2 may seem a bit complicated at first sight, but the underlying ideas are simple. We start by noticing that this algorithm has three important parameters:

1. $a_n \in \{1,\ldots,n\}$: the number of sampled data points in each tree;
2. $m_{\text{try}} \in \{1,\ldots,p\}$: the number of possible directions for splitting at each node of each tree;
3. `nodesize` $\in \{1,\ldots,n\}$: the number of examples in each cell below which the cell is not split.

The algorithm works by growing M different (randomized) trees as follows. Prior to the construction of each tree, a_n observations are drawn at random with replacement from the original data set; then, at each cell of each tree, a split is performed by maximizing the CART-criterion (see below); lastly, construction of individual trees is stopped when each cell contains less than `nodesize` points. By default in the regression mode, the parameter m_{try} is set to $p/3$, a_n is set to n, and `nodesize` is set to 5. The role and influence of these three parameters on the accuracy of the method will be thoroughly discussed in the next section.

We still have to describe how the CART-split criterion operates. With this aim in mind, we let A be a generic cell and denote by $N_n(A)$ the number of data points falling in A. A cut in A is a pair (j,z), where j is some value (dimension) from $\{1,\ldots,p\}$ and z the position of the cut along the j-th coordinate, within the limits of A. Let $\mathcal{C}_A$ be the set of all such possible cuts in A. Then, with the notation $\mathbf{X}_i = (\mathbf{X}_i^{(1)},\ldots,\mathbf{X}_i^{(p)})$, for any $(j,z) \in \mathcal{C}_A$, the CART-split criterion takes the form

Algorithm 2: Breiman's random forest predicted value at $\mathbf{x}$.

Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $a_n \in \{1, \ldots, n\}$, `mtry` $\in \{1, \ldots, p\}$, `nodesize` $\in \{1, \ldots, n\}$, and $\mathbf{x} \in [0,1]^p$.

Output: Prediction of the random forest at $\mathbf{x}$.

for $j = 1, \ldots, M$ **do**
  Select a_n points, with replacement, uniformly in $\mathcal{D}_n$.
  Set $\mathcal{P}_0 = \{[0,1]^p\}$ the partition associated with the root of the tree.
  For all $1 \leq \ell \leq a_n$, set $\mathcal{P}_\ell = \emptyset$.
  Set $n_{nodes} = 1$ and $level = 0$.
  while $n_{nodes} < a_n$ **do**
    if $\mathcal{P}_{level} = \emptyset$ **then**
      $level = level + 1$.
    else
      Let A be the first element in $\mathcal{P}_{level}$.
      if *A contains less than* `nodesize` *points* **then**
        $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$.
        $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A\}$.
      else
        Select uniformly, without replacement, a subset $\mathcal{M}_{\text{try}} \subset \{1, \ldots, p\}$ of cardinality m_{try}.
        Select the best split in A by optimizing the CART-split criterion along the coordinates in $\mathcal{M}_{\text{try}}$ *(see text for details)*.
        Cut the cell A according to the best split. Call A_L and A_R the two resulting cells.
        $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$.
        $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A_L\} \cup \{A_R\}$.
        $n_{nodes} = n_{nodes} + 1$.
      end
    end
  end
  Compute the predicted value $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ at $\mathbf{x}$ equal to the average of the Y_i falling in the cell of $\mathbf{x}$ in partition $\mathcal{P}_{level} \cup \mathcal{P}_{level+1}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n)$ at the query point $\mathbf{x}$ according to (2.1).

$$L_{reg,n}(j,z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \mathbb{1}_{\mathbf{X}_i \in A} - \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z})^2 \mathbb{1}_{\mathbf{X}_i \in A}, \tag{2.2}$$

where $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$, and $\bar{Y}_A$ (resp., $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) is the average of the Y_i belonging to A (resp., A_L, A_R), with the convention $0/0 = 0$. For each cell A, the best cut $(j_n^\star, z_n^\star)$ is selected by maximizing $L_n(j,z)$ over $\mathcal{M}_{try}$ and $\mathcal{C}_A$; that is,

$$(j_n^\star, z_n^\star) \in \underset{\substack{j \in \mathcal{M}_{try} \\ (j,z) \in \mathcal{C}_A}}{\arg\max} L_n(j,z).$$

(To remove some of the ties in the argmax, the best cut is always performed in the middle of two consecutive data points.)

Thus, at each cell of each tree, the algorithm chooses uniformly at random m_{try} coordinates in $\{1, \ldots, p\}$, evaluates criterion (2.2) over all possible cuts in the m_{try} directions, and returns the best one. The quality measure (2.2) is the criterion used in the most influential CART algorithm of Breiman et al. [1984]. This criterion measures the (renormalized) difference between the empirical variance in the node before and after a cut is performed—the only difference here is that it is evaluated over a subset $\mathcal{M}_{\text{try}}$ of randomly selected coordinates, and **not** over the whole range $\{1, \ldots, p\}$. However, contrary to the CART algorithm, the individual trees are not pruned, and the final cells have a cardinality that does not exceed `nodesize`. Also, each tree is constructed on a subset of a_n examples picked within the initial sample, **not** on the whole sample $\mathcal{D}_n$. When $a_n = n$, the algorithm runs in bootstrap mode, whereas $a_n < n$ corresponds to subsampling (with replacement). Last but not least, the process is repeated M (a large number) times.

Classification. In the classification case, if a leaf represents region A, then a randomized tree classifier takes the simple form

$$m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i \in A, Y_i = 1} > \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i \in A, Y_i = 0}, \quad \mathbf{x} \in A \\ 0 & \text{otherwise.} \end{cases}$$

That is, in each leaf, a majority vote is taken over all $(\mathbf{X}_i, Y_i)$ for which $\mathbf{X}_i$ is in the same region. Ties are broken, by convention, in favor of class 0. Algorithm 2 can be easily adapted to do classification by modifying the CART-split criterion for the binary setting. For any cell A, let $p_{0,n}(A)$ (resp., $p_{1,n}(A)$) be the empirical probability that a data point with label 0 (resp. label 1) falls into A. Then, for any $(j,z) \in \mathcal{C}_A$, the classification CART-split criterion takes the form

$$L_{\text{class},n}(j,z) = p_{0,n}(A)p_{1,n}(A) - \frac{N_n(A_L)}{N_n(A)} \times p_{0,n}(A_L)p_{1,n}(A_L) - \frac{N_n(A_R)}{N_n(A)} \times p_{0,n}(A_R)p_{1,n}(A_R). \tag{2.3}$$

This criterion is based on the so-called *Gini impurity measure* $2p_{0,n}(A)p_{1,n}(A)$ [Breiman et al., 1984], which has a simple interpretation. Instead of using the majority vote to classify a data point that falls in cell A, one can use the rule that assigns an observation, selected at random from the node, to label ℓ with probability $p_{\ell,n}(A)$, for $j \in \{0, 1\}$. The estimated probability that the item has actually label ℓ is $p_{\ell,n}(A)$. Therefore the estimated probability of misclassification under this rule is the Gini index $2p_{1,n}(A)p_{2,n}(A)$. When dealing with classification problems, it is usually recommended to set `nodesize` $= 1$ and $m_{\text{try}} = \sqrt{p}$ [see, e.g., Liaw and Wiener, 2002].

2.2.3 Parameter tuning

Literature focusing on tuning the parameters M, `mtry`, `nodesize` and a_n is unfortunately rare, with the notable exception of Díaz-Uriarte and de Andrés [2006], Bernard et al. [2008], and Genuer et al. [2010]. It is easy to see that the forest's variance decreases as M grows. Thus, more accurate predictions are likely to be obtained by choosing a large number of trees. It is interesting to note that picking a large M does not lead to overfitting, since finite forests converge to infinite ones [Breiman, 2001]. However, the computational cost for inducing a forest increases linearly with M, so a good choice results from a trade-off between computational complexity (M should not be too large for the computations to finish in a reasonable time) and accuracy (M must be large enough for predictions to be stable). In this respect, Díaz-Uriarte and de Andrés [2006] argue that the value of M is irrelevant (provided that M is large enough) in a prediction problem involving microarray data sets, where the aim is to classify patients according to their genetic profiles (typically, less than one hundred patients for several thousand genes). For more details we refer the reader to Genuer et al. [2010], who offer a thorough discussion on the choice of this parameter in various regression problems. Another interesting and related approach is by Latinne et al. [2001], who propose a simple procedure that determines *a priori* a minimum number of tree estimates to combine in order to obtain a prediction accuracy level similar to that obtained with a larger forest. Their experimental results show that it is possible to significantly limit the number of trees.

In the `R` package `randomForest`, the default value of the parameter `nodesize` is 1 for classification and 5 for regression. These values are often reported to be good choices [e.g., Díaz-Uriarte and de Andrés, 2006], despite the fact that this is not supported by solid theory. The effect of m_{try} has been thoroughly investigated in Díaz-Uriarte and de Andrés [2006], who show that this parameter has a little impact on the performance of the method, though larger values may be associated with a reduction in the predictive performance. On the other hand, Genuer et al. [2010] claim that the default value of m_{try} is either optimal or too small. Therefore, a conservative approach is to take m_{try} as large as possible (limited by available computing resources) and set $m_{\text{try}} = p$ (recall that p is the dimension of the $\mathbf{X}_i$). A data-driven choice of `mtry` is implemented in the algorithm *Forest-RK* of Bernard et al. [2008].

2.3 Simplified models and local averaging estimates

2.3.1 Simplified models

Despite their widespread use, a gap remains between the theoretical understanding of random forests and their practical performance. This algorithm, which relies on complex data-dependent mechanisms, is difficult to analyze and its basic mathematical properties are still not well understood.

This state of affairs has led to polarization between theoretical and empirical contributions to the literature. Empirically focused papers describe elaborate extensions to the basic random forest framework, adding domain-specific refinements that push the state of the art in performance, but come with no clear guarantees. In contrast, most theoretical papers focus on simplifications or stylized versions of the standard algorithm, where the mathematical analysis is more tractable.

A basic framework to assess the theoretical properties of forests involves models that are calibrated independently of the training set $\mathcal{D}_n$. This family of simplified models is often called *purely random forests*. A widespread example is the *centered forest*, whose principle is as follows: (i) there is no bootstrap step; (ii) at each node of each individual tree, a coordinate is uniformly chosen in $\{1, \ldots, p\}$; and (iii) a split is performed at the center of the cell along the selected coordinate. The operations (ii)-(iii) are recursively repeated k times, where $k \in \mathbb{N}$ is a parameter of the algorithm. The procedure stops when a full binary tree with k levels is reached, so that each tree ends up with exactly 2^k leaves. The parameter k acts as a smoothing parameter that controls the size of the terminal cells (see Figure 2.1 for an example in two dimensions). It should be chosen large enough in order to detect local changes in the distribution, but not too much to guarantee an effective averaging process in the leaves. In *uniform random forests*, a variant of centered forests, cuts are performed uniformly at random over the range of the selected coordinate, not at the center. Modulo some minor modifications, their analysis is similar.

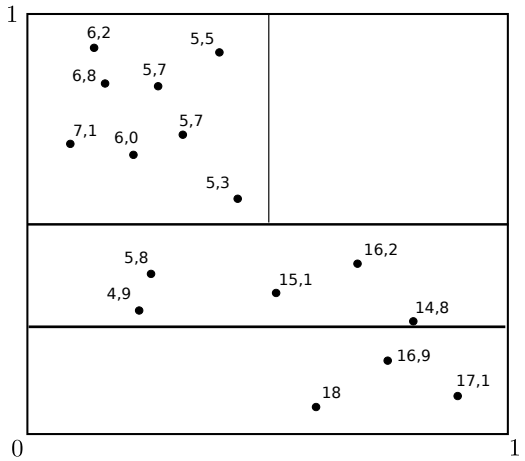

Figure 2.1: A centered tree at level 2.

The centered forest rule was formally analyzed in Biau et al. [2008] and Scornet [2014], who

proved that the method is consistent (both for classification and regression) provided $k \to \infty$ and $n/2^k \to \infty$. The proof relies on a general consistency result for random trees stated in Devroye et al. [1996, Chapter 6]. If $\mathbf{X}$ is uniformly distributed in $[0,1]^p$, then there are on average about $n/2^k$ data points per terminal node. In particular, the choice $k \approx \log n$ corresponds to obtaining a small number of examples in the leaves, in accordance with Breiman's [2001] idea that the individual trees should not be pruned. Unfortunately, this choice of k does not satisfy the condition $n/2^k \to \infty$, so something is lost in the analysis. Moreover, the bagging step is absent, and forest consistency is obtained as a by-product of tree consistency. Overall, this model does not demonstrate the benefit of using forests in place of individual trees and is too simple to explain the mathematical forces driving Breiman's forests.

The rates of convergence of centered forests are discussed in Breiman [2004] and Biau [2012]. In their approaches, the target regression function $m(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$, which is originally a function of $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})$, is assumed to depend only on a nonempty subset $\mathcal{S}$ (for $\mathcal{S}$trong) of the p features. Thus, letting $\mathbf{X}_\mathcal{S} = (X^{(j)} : j \in \mathcal{S})$, we have

$$m(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}_\mathcal{S}].$$

The variables of the remaining set $\{1, \ldots, p\} \backslash \mathcal{S}$ have no influence on the response Y and can be safely removed. In this dimension reduction scenario, the ambient dimension p can be large, much larger than the sample size n, but we believe that the representation is sparse, i.e., that a potentially small number of coordinates of m are active— the ones with indices matching the set $\mathcal{S}$. Letting $|\mathcal{S}|$ be the cardinality of $\mathcal{S}$, the value $|\mathcal{S}|$ characterizes the sparsity of the model: the smaller $|\mathcal{S}|$, the sparser m.

Breiman [2004] and Biau [2012] proved that if the random trees are grown by using coordinates in $\mathcal{S}$ with high probability, and if m satisfies a Lipschitz-type smoothness assumption, then

$$\mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = \mathrm{O}\Big(n^{\frac{-0.75}{|\mathcal{S}| \log 2 + 0.75}}\Big).$$

This equality shows that the rate of convergence of m_n to m depends only on the number $|\mathcal{S}|$ of strong variables, not on the ambient dimension p. This rate is strictly faster than the usual rate $n^{-2/(p+2)}$ as soon as $|\mathcal{S}| \leq \lfloor 0.54p \rfloor$. In effect, the intrinsic dimension of the regression problem is $|\mathcal{S}|$, not p, and we see that the random forest estimate cleverly adapts itself to the sparse framework. This property may be useful for high-dimensional regression, when the number of variables is much larger than the sample size. It may also explain why random forests are able to handle a large number of input variables without overfitting.

An alternative model for pure forests, called *purely uniform random forests* (PURF) is discussed in Genuer [2012]. For $p = 1$, a PURF is obtained by drawing k random variables uniformly on $[0,1]$, and subsequently dividing $[0,1]$ into random sub-intervals. Although this construction is not exactly recursive, it is equivalent to growing a decision tree by deciding at each level which node to split with a probability equal to its length. Genuer [2012] proves that PURF are consistent and, under a Lipschitz assumption, that the estimate satisfies

$$\mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2 = \mathrm{O}\left(n^{-2/3}\right).$$

This rate is minimax over the class of Lipschitz functions [Stone, 1980, 1982].

It is often acknowledged that random forests reduce the estimation error of a single tree, while maintaining the same approximation error. In this respect, Biau [2012] argues that the estimation error of centered forests tends to zero (at the slow rate $1/\log n$) even if each tree is fully grown (i.e., $k \approx \log n$). This result is a consequence of the tree-averaging process, since the estimation error of an individual fully grown tree does not tend to zero. Unfortunately, the choice $k \approx \log n$ is too large to ensure consistency of the corresponding forest, whose approximation error remains constant. Similarly, Genuer [2012] shows that the estimation error of PURF is reduced by a factor of 0.75 compared to the estimation error of individual trees. The most recent attempt to assess the gain of forests in terms of estimation and approximation errors is by Arlot and Genuer [2014], who claim that the rate of the approximation error of certain models is faster than that of the individual trees.

2.3.2 Forests, neighbors and kernels

Let us consider a sequence of independent and identically distributed random variables X_1, $\ldots, X_n$. In random geometry, a random observation $\mathbf{X}_i$ is said to be a *layered nearest neighbor* (LNN) of a point $\mathbf{x}$ (from $\mathbf{X}_1, \ldots, \mathbf{X}_n$) if the hyperrectangle defined by $\mathbf{x}$ and $\mathbf{X}_i$ contains no other data points (Barndorff-Nielsen and Sobel, 1966, Bai et al., 2005; see also Devroye et al., 1996, Chapter 11, Problem 6). As illustrated in Figure 2.2, the number of LNN of $\mathbf{x}$ is typically larger than one and depends on the number and configuration of the sample points.

Surprisingly, the LNN concept is intimately connected to random forests. Indeed, if exactly one point is left in the leaves, then no matter what splitting strategy is used, the forest estimate at $\mathbf{x}$ is but a weighted average of the Y_i whose corresponding $\mathbf{X}_i$ are LNN of $\mathbf{x}$. In other words,

$$m_{\infty,n}(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) Y_i, \tag{2.4}$$

where the weights $(W_{n1}, \ldots, W_{nn})$ are nonnegative functions of the sample $\mathcal{D}_n$ that satisfy $W_{ni}(\mathbf{x}) = 0$ if $\mathbf{X}_i$ is not an LNN of $\mathbf{x}$ and $\sum_{i=1}^{n} W_{ni} = 1$. This important connection was first pointed out by Lin and Jeon [2006], who proved that if $\mathbf{X}$ is uniformly distributed on $[0,1]^p$ then, provided tree growing is independent of $Y_1, \ldots, Y_n$ (such simplified models are sometimes called *non-adaptive*), we have

$$\mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = \mathrm{O}\left(\frac{1}{n_{\max}(\log n)^{p-1}}\right),$$

where $n_{\max}$ is the maximal number of points in the terminal cells (Biau and Devroye, 2010, extended this inequality to the case where $\mathbf{X}$ has a density on $[0,1]^p$). Unfortunately, the exact values of the weight vector $(W_{n1}, \ldots, W_{nn})$ attached to the original random forest algorithm are unknown, and a general theory of forests in the LNN framework is still undeveloped.

It remains however that equation (2.4) opens the way to the analysis of random forests via a local-averaging approach, i.e., via the average of those Y_i for which $\mathbf{X}_i$ is "close" to $\mathbf{x}$ [Györfi et al., 2002]. Indeed, observe, starting from (2.1), that for a finite forest with M trees, we have

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} \frac{Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}}{N_n(\mathbf{x}, \Theta_j)} \right),$$

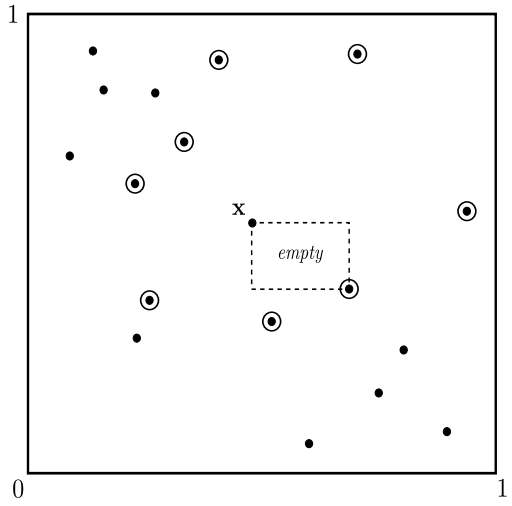

Figure 2.2: The layered nearest neighbors (LNN) of a point $\mathbf{x}$ in dimension $p = 2$.

where $A_n(\mathbf{x}, \Theta_j)$ is the cell containing $\mathbf{x}$ and $N_n(\mathbf{x}, \Theta_j) = \sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}$ is the number of data points falling in $A_n(\mathbf{x}, \Theta_j)$. Thus,

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M) = \sum_{i=1}^n W_{ni}(\mathbf{x}) Y_i,$$

where the weights $W_{ni}(\mathbf{x})$ are defined by

$$W_{ni}(\mathbf{x}) = \frac{1}{M} \sum_{j=1}^M \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}}{N_n(\mathbf{x}, \Theta_j)}.$$

It is easy to see that the W_{ni} are nonnegative and sum to one if the cell containing $\mathbf{x}$ is not empty. Thus, the contribution of observations falling into cells with a high density of data points is smaller than the contribution of observations belonging to less-populated cells. This remark is especially true when the forests are built independently of the data set—for example, PURF—since, in this case, the number of examples in each cell is not controlled. Next, if we let M tend to infinity, then the estimate $m_{\infty,n}$ may be written (up to some negligible terms)

$$m_{\infty,n}(\mathbf{x}) \approx \frac{\sum_{i=1}^n Y_i K_n(\mathbf{X}_i, \mathbf{x})}{\sum_{j=1}^n K_n(\mathbf{X}_j, \mathbf{x})}, \tag{2.5}$$

where

$$K_n(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta \left[\mathbf{z} \in A_n(\mathbf{x}, \Theta) \right].$$

The function $K_n(\cdot, \cdot)$ is called the *kernel* and characterizes the shape of the "cells" of the infinite random forest. The quantity $K_n(\mathbf{x}, \mathbf{z})$ is nothing but the probability that $\mathbf{x}$ and $\mathbf{z}$ are connected (i.e., they fall in the same cell) in a random tree. Therefore, the kernel K_n can be seen as

a proximity measure between two points in the forest. Hence, any forest has its own metric K_n, but unfortunately the one associated with Breiman's forest is strongly data-dependent and therefore complicated to work with.

It should be noted that K_n does not necessarily belong to the family of Nadaraya-Watson-type kernels [Nadaraya, 1964, Watson, 1964], which satisfy a homogeneous property of the form $K_h(\mathbf{x},\mathbf{z}) = \frac{1}{h}K((\mathbf{x}-\mathbf{z})/h)$ for some *smoothing parameter* $h > 0$. The analysis of estimates of the form (2.5) is, in general, more complicated, depending of the type of forest under investigation. For example, Scornet [2015] proved that for a centered forest defined on $[0,1]^p$ with parameter k, we have

$$K_{k,n}(\mathbf{x},\mathbf{z}) = \sum_{\substack{k_1,\ldots,k_p \\ \sum_{j=1}^p k_j = k}} \frac{k!}{k_1!\ldots k_p!}\left(\frac{1}{p}\right)^k \prod_{j=1}^p \mathbb{1}_{\lceil 2^{k_j} x_j\rceil = \lceil 2^{k_j} z_j\rceil}$$

($\lceil\cdot\rceil$ is the ceiling function). As an illustration, Figure 2.3 shows the graphical representation for $k = 1$, 2 and 5 of the function f_k defined by

$$\begin{array}{rccl} f_k: & [0,1]\times[0,1] & \to & [0,1] \\ & \mathbf{z} = (z_1, z_2) & \mapsto & K_{k,n}\big((\frac{1}{2},\frac{1}{2}),\mathbf{z}\big). \end{array}$$

The connection between forests and kernel estimates is mentioned in Breiman [2000a] and developed in detail in Geurts et al. [2006]. The most recent advances in this direction are by Arlot and Genuer [2014], who show that a simplified forest model can be written as a kernel estimate, and provide its rates of convergence. On the practical side, Davies and Ghahramani [2014] highlight the fact that using Gaussian processes with a specific kernel-based random forest can empirically outperform state-of-the-art Gaussian process methods. Besides, kernel-based random forests can be used as the input for a large variety of existing kernel-type methods such as Kernel Principal Component Analysis and Support Vector Machines.

2.4 Theory for Breiman's forests

This section deals with Breiman's [2001] original algorithm. Since the construction of Breiman's forests depends on the whole sample $\mathcal{D}_n$, a mathematical analysis of the whole algorithm is difficult. To move forward, the individual mechanisms at work in the procedure have been investigated separately, namely the resampling step and the splitting scheme.

2.4.1 The resampling mechanism

The resampling step in Breiman's [2001] original algorithm is performed by choosing n times from of n points with replacement to grow the individual trees. This procedure, which traces back to the work of Efron [1982] [see also Politis et al., 1999], is called the *bootstrap* in the statistical literature. The idea of generating many bootstrap samples and averaging predictors is called *bagging* (bootstrap-aggregating). It was suggested by Breiman [1996] as a simple way to improve the performance of weak or unstable learners. Although one of the great advantages of the bootstrap is its simplicity, the theory turns out to be complex. In effect, the bootstrapped observations have a distribution that is different from the original one, as the following example

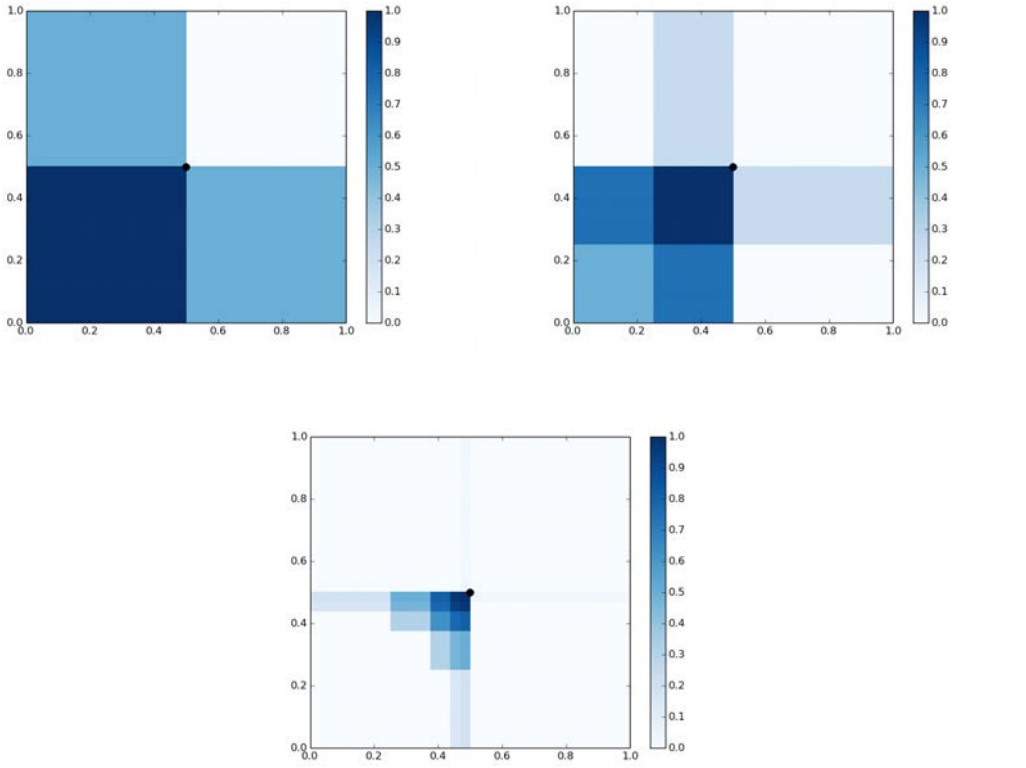

Figure 2.3: Representations of f_1, f_2 and f_5 in $[0,1]^2$.

shows. Assume that $\mathbf{X}$ has a density, and note that whenever the data points are sampled with replacement, then with positive probability, at least one observation from the original sample will be selected more than once. Therefore, the resulting $\mathbf{X}_i$ of the bootstrapped sample cannot have an absolutely continuous distribution.

The role of the bootstrap in random forests is still poorly understood and, to date, most analyses are doomed to replace the bootstrap by a subsampling scheme, assuming that each tree is grown with $a_n < n$ examples randomly chosen without replacement from the initial sample [Mentch and Hooker, 2014a, Wager, 2014, Scornet et al., 2015b]. Most of the time, the subsampling rate a_n/n is assumed to tend to zero at some prescribed rate—an assumption that excludes *de facto* the bootstrap regime. In this respect, the analysis of so-called *median random forests* by Scornet [2014] provides some insight as to the role and importance of subsampling. The assumption $a_n/n \to 0$ guarantees that every single observation pair $(\mathbf{X}_i, Y_i)$ is used in the m-th tree's construction with a probability that becomes small as n grows. It also forces the query point $\mathbf{x}$ to be disconnected from $(\mathbf{X}_i, Y_i)$ in a large proportion of trees. Indeed, if this were not the case, then the predicted value at $\mathbf{x}$ would be overly influenced by the single pair $(\mathbf{X}_i, Y_i)$, which would make the ensemble inconsistent. In fact, the estimation error of the median forest estimate is small as soon as the maximum probability of connection between the query point

and all observations is small. Thus, the assumption $a_n/n \to 0$ is but a convenient way to control these probabilities, by ensuring that partitions are dissimilar enough.

Biau and Devroye [2010] noticed that Breiman's bagging principle has a simple application in the context of nearest neighbor methods. Recall that the 1-nearest neighbor (1-NN) regression estimate sets $r_n(\mathbf{x}) = Y_{(1)}(\mathbf{x})$, where $Y_{(1)}(\mathbf{x})$ corresponds to the feature vector $\mathbf{X}_{(1)}(\mathbf{x})$ whose Euclidean distance to $\mathbf{x}$ is minimal among all $\mathbf{X}_1, \ldots, \mathbf{X}_n$. (Ties are broken in favor of smallest indices.) It is clearly not, in general, a consistent estimate [Devroye et al., 1996, Chapter 5]. However, by bagging, one may turn the 1-NN estimate into a consistent one, provided that the size of resamples is sufficiently small. We proceed as follows, via a randomized basic regression estimate r_{a_n} in which $1 \leq a_n \leq n$ is a parameter. The elementary predictor r_{a_n} is the 1-NN rule for a random subsample of size a_n drawn with (or without) replacement from $\mathcal{D}_n$. We apply bagging, that is, we repeat the random sampling an infinite number of times and take the average of the individual outcomes. Thus, the bagged regression estimate $r_n^\star$ is defined by

$$r_n^\star(\mathbf{x}) = \mathbb{E}^\star \left[r_{a_n}(\mathbf{x}) \right],$$

where $\mathbb{E}^\star$ denotes expectation with respect to the resampling distribution, conditional on the data set $\mathcal{D}_n$. Biau and Devroye [2010] proved that the estimate $r_n^\star$ is universally (i.e., without conditions on the distribution of $(\mathbf{X}, Y)$) mean squared consistent, provided $a_n \to \infty$ and $a_n/n \to 0$. The proof relies on the observation that $r_n^\star$ is in fact a weighted nearest neighbor estimate [Stone, 1977] with weights

$$W_{ni} = \mathbb{P}(i\text{-th nearest neighbor of } \mathbf{x} \text{ is the 1-NN in a random selection}).$$

The connection between bagging and nearest neighbor estimation is further explored by Biau et al. [2010], who prove that the bagged estimate $r_n^\star$ achieves optimal rate of convergence over Lipschitz smoothness classes, independently from the fact that resampling is done with or without replacement.

2.4.2 Decision splits

The coordinate-split process of the random forest algorithm is not easy to grasp, essentially because it uses both the $\mathbf{X}_i$ and Y_i variables to make its decision. Building upon the ideas of Bühlmann and Yu [2002], Banerjee and McKeague [2007] establish a limit law for the split location in the context of a regression model of the form $Y = m(\mathbf{X}) + \varepsilon$, where $\mathbf{X}$ is real-valued and ε an independent Gaussian noise. In essence, their result is as follows. Assume for now that the distribution of $(\mathbf{X}, Y)$ is known, and denote by $d^\star$ the (optimal) split that maximizes the theoretical CART-criterion at a given node. In this framework, the regression estimates restricted to the left and right children of the cell takes the respective forms

$$\beta_{\ell,n}^\star = \mathbb{E}[Y | X \leq d^\star] \quad \text{and} \quad \beta_{r,n}^\star = \mathbb{E}[Y | X > d^\star].$$

When the distribution of $(\mathbf{X}, Y)$ is unknown, so are $\beta_\ell^\star$, $\beta_r^\star$ and $d^\star$, and these quantities are estimated by their natural empirical counterparts:

$$(\hat{\beta}_{\ell,n}, \hat{\beta}_{r,n}, \hat{d}_n) \in \operatorname*{arg\,min}_{\beta_\ell, \beta_r, d} \sum_{i=1}^n \left[Y_i - \beta_\ell \mathbb{1}_{X_i \leq d} - \beta_r \mathbb{1}_{X_i > d} \right]^2.$$

Assuming that the model satisfies some regularity assumptions (in particular, $\mathbf{X}$ has a density f, and both f and m are continuously differentiable), Banerjee and McKeague [2007] prove that

$$n^{1/3}(\hat{\beta}_{\ell,n} - \beta_\ell^\star, \hat{\beta}_{r,n} - \beta_r^\star, \hat{d}_n - d^\star) \xrightarrow{\mathcal{D}} (c_1, c_2, 1) \arg\max_t Q(t), \tag{2.6}$$

where $\mathcal{D}$ denotes convergence in distribution, $Q(t) = aW(t) - bt^2$, and W is a standard two-sided Brownian motion process on the real line. Both a and b are positive constants that depend upon the model parameters and the unknown quantities $\beta_\ell^\star$, $\beta_r^\star$ and $d^\star$. The limiting distribution in (2.6) allows one to construct confidence intervals for the position of CART-splits. Interestingly, Banerjee and McKeague [2007] refer to the study of Qian et al. [2003] on the effects of phosphorus pollution in the Everglades, which uses split points in a novel way. There, the authors identify threshold levels of phosphorus concentration that are associated with declines in the abundance of certain species. In their approach, split points are not just a means to build trees and forests, but can also provide important information on the structure of the underlying distribution.

A further analysis of the behavior of forest splits is performed by Ishwaran [2013], who argues that the so-called *end-cut preference* (ECP) of the CART-splitting procedure (that is, the fact that splits along non-informative variables are likely to be near the edges of the cell—see Breiman et al., 1984) can be seen as a desirable property. Given the randomization mechanism at work in forests, there is indeed a positive probability that none of the preselected variables at a node are informative. When this happens, and if the cut is performed, say, at the center of a side of the cell, then the sample size of the two resulting cells is drastically reduced by a factor of two—this is an undesirable property, which may be harmful for the prediction task. In other words, Ishwaran [2013] stresses that the ECP property ensures that a split along a noisy variable is performed near the edge, thus maximizing the tree node sample size and making it possible for the tree to recover from the split downstream. Ishwaran [2013] claims that this property can be of benefit even when considering a split on an informative variable, if the corresponding region of space contains little signal.

There exists a variety of random forest variants based on the CART-criterion. For example, the *Extra-Tree* algorithm of Geurts et al. [2006] consists in randomly selecting a set of split points and then choosing the split that maximizes the CART-criterion. This algorithm has similar accuracy performance while being more computationally efficient. In the *PERT* (Perfect Ensemble Random Trees) approach of Cutler and Zhao [2001], one builds perfect-fit classification trees with random split selection. While individual trees clearly overfit, the authors claim that the whole procedure is eventually consistent since all classifiers are believed to be almost uncorrelated. Let us also mention that additional randomness can be added in the tree construction by considering splits along linear combinations of features. This idea, due to Breiman [2001], has been implemented by Truong [2009] in the package `obliquetree` of statistical computing environment `R`.

2.4.3 Asymptotic normality and consistency

All in all, little has been proven mathematically for the original procedure of Breiman [2001]. Recently, consistency and asymptotic normality of the whole algorithm were proved under simplifications of the procedure (replacing bootstrap by subsampling and simplifying the splitting

step). Wager [2014] proves the asymptotic normality of the method and establishes that the infinitesimal jackknife consistently estimates the forest variance. A similar result on the asymptotic normality of finite forests, proved by Mentch and Hooker [2014a], states that whenever M (the number of trees) is allowed to vary with n, and when $a_n = o(\sqrt{n})$ and $\lim_{n\to\infty} n/M_n = 0$, then for a fixed $\mathbf{x}$,

$$\frac{\sqrt{n}(m_{M,n}(\mathbf{x};\Theta_1,\ldots,\Theta_M) - m_{\infty,n}(\mathbf{x}))}{\sqrt{a_n^2\zeta_{1,a_n}}} \xrightarrow{\mathcal{D}} N,$$

where N is a standard normal random variable,

$$\zeta_{1,a_n} = \text{Cov}\left[m_n(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_{a_n};\Theta), m_n(\mathbf{X}_1,\mathbf{X}_2',\ldots,\mathbf{X}_{a_n}';\Theta')\right],$$

$\mathbf{X}_i'$ an independent copy of $\mathbf{X}_i$ and Θ' an independent copy of Θ. Note that in this model, both the sample size and the number of trees grow to infinity. Recently, Scornet et al. [2015b] proved a consistency result in the context of additive regression models for the pruned version of Breiman's forest. Unfortunately, the consistency of the unpruned procedure comes at the price of a conjecture regarding the behavior of the CART algorithm that is difficult to verify.

We close this section with a negative but interesting result due to Biau et al. [2008]. In this example, the total number k of cuts is fixed and `mtry` $= 1$. Furthermore, each tree is built by minimizing the true probability of error at each node. Consider the joint distribution of $(\mathbf{X}, Y)$ sketched in Figure 2.4 and let $m(\mathbf{x}) = \mathbb{P}[Y = 1|\mathbf{X} = \mathbf{x}]$. The variable $\mathbf{X}$ has a uniform distribution on $[0,1]^2 \cup [1,2]^2 \cup [2,3]^2$ and Y is a function of $\mathbf{X}$—that is, $m(\mathbf{x}) \in \{0,1\}$ and $L^* = 0$—defined as follows. The lower left square $[0,1] \times [0,1]$ is divided into countably infinitely many vertical strips in which the strips with $m(\mathbf{x}) = 0$ and $m(\mathbf{x}) = 1$ alternate. The upper right square $[2,3] \times [2,3]$ is divided similarly into horizontal strips. The middle rectangle $[1,2] \times [1,2]$ is a 2×2 checkerboard. It is easy to see that no matter what the sequence of random selection of split directions is and no matter for how long each tree is grown, no tree will ever cut the middle rectangle and therefore the probability of error of the corresponding random forest classifier is at least $1/6$. This example illustrates that consistency of greedily grown random forests is a delicate issue. Note however that if Breiman's [2001] original algorithm is used in this example (i.e., when all cells with more than one data point in them are split) then one obtains a consistent classification rule.

2.5 Variable selection

2.5.1 Variable importance measures

Random forests can be used to rank the importance of variables in regression or classification problems via two measures of significance. The first, called *Mean Decrease Impurity* (MDI), is based on the total decrease in node impurity from splitting on the variable, averaged over all trees. The second, referred to as *Mean Decrease Accuracy* (MDA), stems from the idea that if the variable is not important, then rearranging its values should not degrade prediction accuracy.

Set $\mathbf{X} = (\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(p)})$. For a forest resulting from the aggregation of M trees, the MDI

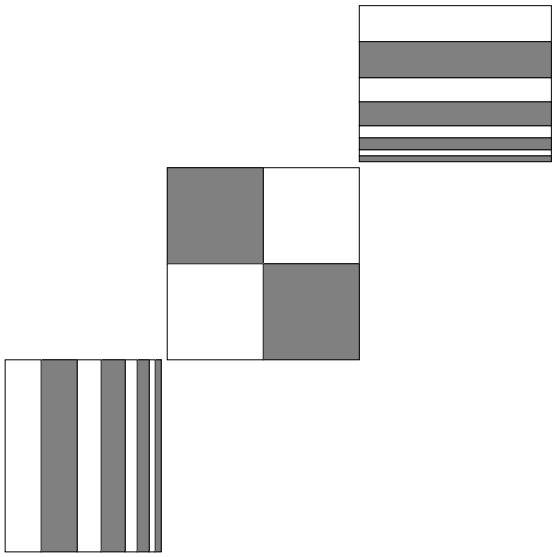

Figure 2.4: An example of a distribution for which greedy random forests are inconsistent. The distribution of $\mathbf{X}$ is uniform on the union of the three large squares. White areas represent the set where $m(\mathbf{x}) = 0$ and grey where $m(\mathbf{x}) = 1$.

of the variable $X^{(j)}$ is defined as

$$\widehat{\mathrm{MDI}}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \sum_{\substack{t \in \mathcal{T}_\ell \\ j^\star_{n,t}=j}} 2 p_{n,t} L_{\mathrm{reg},n}(j^\star_{n,t}, z^\star_{n,t}),$$

where $p_{n,t}(t)$ is the fraction of observations falling in the node t, $\{\mathcal{T}_\ell\}_{1 \leq \ell \leq M}$ the collection of trees in the forest, and $(j^\star_{n,t}, z^\star_{n,t})$ the split that maximizes the empirical criterion (2.2) in node t. Note that the same formula holds for classification random forests by replacing the criterion $L_{\mathrm{reg},n}$ by its classification counterpart $L_{\mathrm{class},n}$. Thus, the MDI of $X^{(j)}$ computes the weighted decrease of impurity corresponding to splits along the variable $X^{(j)}$ and averages this quantity over all trees.

The MDA relies on a different principle and uses the so-called *out-of-bag* error estimate. In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test error. It is estimated internally, during the run, as follows. Since each tree is constructed using a different bootstrap sample from the original data, about one-third of cases are left out of the bootstrap sample and not used in the construction of the m-th tree. In this way, for each tree, a test set—disjoint from the training set— is obtained, and averaging over all these left-out cases and over all trees is known as the out-of-bag error estimate.

To measure the importance of the j-th feature, we randomly permute the values of variable $X^{(j)}$ in the out-of-bag cases and put these cases down the tree. The MDA of $X^{(j)}$ is obtained by averaging the difference in out-of-bag error estimation before and after the permutation over all trees. In mathematical terms, consider a variable $X^{(j)}$ and denote by $\mathcal{D}_{\ell,n}$ the out-of-bag

test of the ℓ-th tree and $\mathcal{D}_{\ell,n}^j$ the same data set where the values of $X^{(j)}$ have been randomly permuted. Recall that $m_n(\cdot, \Theta_\ell)$ stands for the ℓ-th tree estimate. Then, by definition,

$$\widehat{\mathrm{MDA}}(X^{(j)}) = \frac{1}{M}\sum_{\ell=1}^{M}\Big[R_n\big[m_n(\cdot,\Theta_\ell),\mathcal{D}_{\ell,n}^j\big] - R_n\big[m_n(\cdot,\Theta_\ell),\mathcal{D}_{\ell,n}\big]\Big], \tag{2.7}$$

where R_n is defined for $\mathcal{D} = \mathcal{D}_{\ell,n}$ or $\mathcal{D} = \mathcal{D}_{\ell,n}^j$ by

$$R_n\big[m_n(\cdot,\Theta_\ell),\mathcal{D}\big] = \frac{1}{|\mathcal{D}|}\sum_{i:(\mathbf{X}_i,Y_i)\in\mathcal{D}}(Y_i - m_n(\mathbf{X}_i,\Theta_\ell))^2.$$

It is easy to see that the population version of $\widehat{\mathrm{MDA}}(X^{(j)})$ takes the form

$$\mathrm{MDA}^\star(X^{(j)}) = \mathbb{E}\big[Y - m_n(\mathbf{X}_j',\Theta)\big]^2 - \mathbb{E}\big[Y - m_n(\mathbf{X},\Theta)\big]^2,$$

where $\mathbf{X}_j' = (X^{(1)},\ldots,X'^{(j)},\ldots,X^{(p)})$ and $X'^{(j)}$ is an independent copy of $X^{(j)}$. For classification purposes, the MDA still satisfies (2.7) with $R_n(m_n(\cdot,\Theta)\ ,\mathcal{D})$ the number of points that are correctly classified by $m_n(\cdot,\Theta)$ in $\mathcal{D}$.

2.5.2 Theoretical results

In the context of a pair of categorical variables $(\mathbf{X}, Y)$, where $\mathbf{X}$ takes finitely many values in, say, $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, Louppe et al. [2013] consider totally randomized and fully developed trees. At each cell, the ℓ-th tree is grown by selecting a variable $X^{(j)}$ uniformly among the features that have not been used in the parent nodes, and by subsequently dividing the cell into $|\mathcal{X}_j|$ children (so the number of children equals the number of modalities of the selected variable). In this framework, it can be shown that the population version of $\mathrm{MDI}(X^{(j)})$ for a single tree satisfies

$$\mathrm{MDI}^\star(X^{(j)}) = \sum_{k=0}^{p-1}\frac{1}{\binom{k}{p}(p-k)}\sum_{B\in\mathcal{P}_k(V^{-j})} I(X_j;Y|B),$$

where $V^{-j} = \{1,\ldots,j-1,j+1,\ldots,p\}$, $\mathcal{P}_k(V^{-j})$ the set of subsets of V^{-j} of cardinality k, and $I(X^{(j)};Y|B)$ the *conditional mutual information* of $X^{(j)}$ and Y given the variables in B. In addition,

$$\sum_{j=1}^{p}\mathrm{MDI}^\star(X^{(j)}) = I(X^{(1)},\ldots,X^{(p)};Y).$$

These results show that the information $I(X^{(1)},\ldots,X^{(p)};Y)$ is the sum of the importances of each variable, which can itself be made explicit using the information values $I(X^{(j)};Y|B)$ between each variable $X^{(j)}$ and the output Y, conditional on variable subsets B of different sizes.

Louppe et al. [2013] define a variable $X^{(j)}$ as irrelevant with respect to $B \subset V = \mathcal{X}_1\times\cdots\times\mathcal{X}_p$ whenever $I(X^{(j)};Y|B) = 0$. Thus, $X^{(j)}$ is irrelevant if and only if $\mathrm{MDI}^\star(X^{(j)}) = 0$. It is easy

to see that if an additional irrelevant variable $X^{(p+1)}$ is added to the list of variables, then the variable importance of any of the $X^{(j)}$ computed with a single tree does not change if the tree is built with the new collection of variables $V \cup \{X^{(p+1)}\}$. In other words, building a tree with an additional irrelevant variable does not change the importances of the other variables.

The most notable results regarding MDA are due to Ishwaran [2007], who studies a slight modification of the criterion via feature noising. To add noise to a variable $X^{(j)}$, one considers a new observation $\mathbf{X}$, take $\mathbf{X}$ down the tree and stop when a split is made according to the variable $X^{(j)}$. Then the right or left child node is selected with probability 1/2, and this procedure is repeated for each subsequent node (whether it is performed along the variable $X^{(j)}$ or not). The importance of variable $X^{(j)}$ is still computed by comparing the error of the forest with that of the "noisy" forest. Assuming that the forest is consistent and that the regression function is piecewise constant, Ishwaran [2007] gives the asymptotic behavior of $\widehat{\mathrm{MDA}}(X^{(j)})$ when the sample size tends to infinity. This behavior is intimately related to the set of subtrees (of the initial regression tree) whose roots are split along the coordinate $X^{(j)}$.

Let us lastly mention the approach of Gregorutti et al. [2013], who computed the MDA criterion for several distributions of $(\mathbf{X}, Y)$. For example, consider a model of the form

$$Y = m(\mathbf{X}) + \varepsilon,$$

where $(\mathbf{X}, \varepsilon)$ is a Gaussian random vector, and assume that the correlation matrix C satisfies $C = [\mathrm{Cov}(X_j, X_k)]_{1 \leq j,k \leq p} = (1-c)I_p + c\mathbb{1}\mathbb{1}^\top$ (the symbol $\top$ denotes transposition, $\mathbb{1} = (1, \ldots, 1)^\top$, and c is a constant in $(0, 1)$). Assume, in addition, that $\mathrm{Cov}(X_j, Y) = \tau_0$ for all $j \in \{1, \ldots, p\}$. Then, for all j,

$$\mathrm{MDI}^\star(X^{(j)}) = 2\left(\frac{\tau_0}{1 - c + pc}\right)^2.$$

Thus, in the Gaussian setting, the variable importance decreases as the inverse of the square of p when the number of correlated variables p increases.

2.5.3 Related works

The empirical properties of the MDA criterion have been extensively analyzed and compared in the statistical computing literature. Indeed, Archer and Kimes [2008], Strobl et al. [2008], Nicodemus and Malley [2009], Auret and Aldrich [2011], and Tolo̧si and Lengauer [2011] stress the negative effect of correlated variables on MDA performance. In this respect, Genuer et al. [2010] noticed that MDA is less able to detect the most relevant variables when the number of correlated features increases. Similarly, the empirical study of Archer and Kimes [2008] points out that both MDA and MDI behave poorly when correlation increases—these results have been experimentally confirmed by Auret and Aldrich [2011] and Tolo̧si and Lengauer [2011]. An argument of Strobl et al. [2008] to justify the bias of MDA in the presence of correlated variables is that the algorithm evaluates the marginal importance of the variables instead of taking into account their effect conditional on each other. A way to circumvent this issue is to combine random forests and the *Recursive Feature Elimination* algorithm of Guyon et al. [2002], as in Gregorutti et al. [2013]. Detecting relevant features can also be achieved via hypothesis testing [Mentch and Hooker, 2014a]—a principle that may be used to detect more complex structures of the regression function, like for instance its additivity [Mentch and Hooker, 2014b].

As for the tree building process, selecting uniformly at each cell a set of features for splitting is simple and convenient, but such procedures inevitably select irrelevant variables. Therefore, several authors have proposed modified versions of the algorithm that incorporate a data-driven weighing of variables. For example, Kyrillidis and Zouzias [2014] study the effectiveness of non-uniform randomized feature selection in decision tree classification, and experimentally show that such an approach may be more effective compared to naive uniform feature selection. *Enriched random forests*, designed by Amaratunga et al. [2008] choose at each node the eligible subsets by weighted random sampling with the weights tilted in favor of informative features. Similarly, the *reinforcement learning trees* (RLT) of Zhu et al. [2012] build at each node a random forest to determine the variable that brings the greatest future improvement in later splits, rather than choosing the one with largest marginal effect from the immediate split.

Choosing weights can also be done via regularization. Deng and Runger [2012] propose a *Regularized Random Forest* (RRF), which penalizes selecting a new feature for splitting when its gain is similar to the features used in previous splits. Deng and Runger [2013] suggest a *Guided RRF* (GRRF), in which the importance scores from an ordinary random forest are used to guide the feature selection process in RRF. Lastly, a Garrote-style convex penalty, proposed by Meinshausen [2009], selects functional groups of nodes in trees, yielding to parcimonious estimates. We also mention the work of Konukoglu and Ganz [2014] who address the problem of controlling the false positive rate of random forests and present a principled way to determine thresholds for the selection of relevant features without any additional computational load.

2.6 Extensions

Weighted forests. In Breiman's [2001] forests, the final prediction is the average of the individual tree outcomes. A natural way to improve the method is to incorporate tree-level weights to emphasize more accurate trees in prediction [Winham et al., 2013]. A closely related idea, proposed by Bernard et al. [2012], is to guide tree building—via resampling of the training set and other *ad hoc* randomization procedures—so that each tree will complement as much as possible the existing trees in the ensemble. The resulting *Dynamic Random Forest* (DRF) shows significant improvement in terms of accuracy on 20 real-based data sets compared to the standard, static, algorithm.

Online forests. In its original version, random forests is an *offline algorithm*, which is given the whole data set from the beginning and required to output an answer. In contrast, *online algorithms* do not require that the entire training set is accessible at once. These models are appropriate for streaming settings, where training data is generated over time and must be incorporated into the model as quickly as possible. Random forests have been extended to the online framework in several ways [Saffari et al., 2009, Denil et al., 2013, Lakshminarayanan et al., 2014]. In Lakshminarayanan et al. [2014], so-called *Mondrian forests* are grown in an online fashion and achieve competitive predictive performance comparable with other online random forests while being faster. When building online forests, a major difficulty is to decide when the amount of data is sufficient to cut a cell. Exploring this idea, Yi et al. [2012] propose *Information Forests*, whose construction consists in deferring classification until a measure of *classification confidence* is sufficiently high, and in fact break down the data so as to maximize

this measure. An interesting theory related to these greedy trees can be found in Biau and Devroye [2013].

Survival forests. Survival analysis attempts to deal with incomplete data, and particularly right-censored data in fields such as clinical trials. In this context, parametric approaches such as proportional hazards are commonly used, but fail to model nonlinear effects. Random forests have been extended to the survival context by Ishwaran et al. [2008], who prove consistency of *Random Survival Forests* (RSF) algorithm assuming that all variables are factors. Yang et al. [2010] showed that by incorporating kernel functions into RSF, their algorithm KIRSF achieves better results in many situations. Ishwaran et al. [2011] review the use of the *minimal depth*, which measures the predictive quality of variables in survival trees.

Ranking forests. Clémençon et al. [2013] have extended random forests to deal with ranking problems and propose an algorithm called *Ranking Forests* based on the ranking trees of Clémençon and Vayatis [2009]. Their approach is based on nonparametric scoring and ROC curve optimization in the sense of the AUC criterion.

Clustering forests. Yan et al. [2013] present a new clustering ensemble method called *Cluster Forests* (CF) in the context of unsupervised classification. CF randomly probes a high-dimensional data cloud to obtain good local clusterings, then aggregates via spectral clustering to obtain cluster assignments for the whole data set. The search for good local clusterings is guided by a cluster quality measure, and CF progressively improves each local clustering in a fashion that resembles tree growth in random forests.

Quantile forests. Meinshausen [2006] shows that random forests provide information about the full conditional distribution of the response variable, and thus can be used for quantile estimation.

Missing data. One of the strengths of random forests is that they can handle missing data. The procedure, explained in Breiman [2003], takes advantage of the so-called *proximity matrix*, which measures the proximity between pairs of observations in the forest, to estimate missing values. This measure is the empirical counterpart of the kernels defined in Section 3.2. Data imputation based on random forests has further been explored by Rieger et al. [2010], Crookston and Finley [2008], and extended to unsupervised classification by Ishioka [2013].

Forests and machine learning. One-class classification is a binary classification task for which only one class of samples is available for learning. Désir et al. [2013] study the *One Class Random Forests* algorithm, which is designed to solve this particular problem. Geremia et al. [2013] have introduced a supervised learning algorithm called *Spatially Adaptive Random Forests* to deal with semantic image segmentation applied to medical imaging protocols. Lastly, in the context of multi-label classification, Joly et al. [2014] adapt the idea of random projections applied to the output space to enhance tree-based ensemble methods by improving accuracy while significantly reducing the computational burden.

Chapter 3

On the asymptotics of random forests

Abstract The last decade has witnessed a growing interest in random forest models which are recognized to exhibit good practical performance, especially in high-dimensional settings. On the theoretical side, however, their predictive power remains largely unexplained, thereby creating a gap between theory and practice. In this paper, we present some asymptotic results on random forests in a regression framework. Firstly, we provide theoretical guarantees to link finite forests used in practice (with a finite number M of trees) to their asymptotic counterparts (with $M = \infty$). Using empirical process theory, we prove a uniform central limit theorem for a large class of random forest estimates, which holds in particular for Breiman's [2001] original forests. Secondly, we show that infinite forest consistency implies finite forest consistency and thus, we state the consistency of several infinite forests. In particular, we prove that q quantile forests—close in spirit to Breiman's [2001] forests but easier to study—are able to combine inconsistent trees to obtain a final consistent prediction, thus highlighting the benefits of random forests compared to single trees.

We would like to thank the two referees for valuable comments and insightful suggestions and Luc Devroye for a substantial improvement of the proof of Lemma 3.2.

Contents

3.1 Introduction

Random forests are a class of algorithms used to solve classification and regression problems. As ensemble methods, they grow several trees as base estimates and aggregate them to make a prediction. In order to obtain many different trees based on a single training set, random forests procedures introduce randomness in the tree construction. For instance, trees can be built by randomizing the set of features [Dietterich and Kong, 1995, Ho, 1998], the data set [Breiman, 1996, 2000b], or both at the same time [Breiman, 2001, Cutler and Zhao, 2001].

Among all random forest algorithms, the most popular one is that of Breiman [2001], which relies on CART procedure [Classification and Regression Trees, Breiman et al., 1984] to grow the individual trees. As highlighted by several applied studies [see, e.g., Hamza and Laroque, 2005, Díaz-Uriarte and de Andrés, 2006], Breiman's [2001] random forests can outperform state-of-the-art methods. They are recognized for their ability to handle high-dimensional data sets, thus being useful in fields such as genomics [Qi, 2012] and pattern recognition [Rogez et al., 2008], just to name a few. On the computational side, Breiman's [2001] forests are easy to run and robust to changes in the parameters they depend on [Liaw and Wiener, 2002, Genuer et al., 2008]. Besides, extensions have been developed in ranking problems [Clémençon et al., 2013], quantile estimation [Meinshausen, 2006], and survival analysis [Ishwaran et al., 2008]. Interesting new developments in the context of massive data sets have been achieved. For instance, Geurts et al. [2006] modified the procedure to reduce calculation time, while other authors extended the procedure to online settings [Denil et al., 2013, Lakshminarayanan et al., 2014, and the references therein].

While Breiman's [2001] forests are extensively used in practice, some of their mathematical properties remain under active investigation. In fact, most theoretical studies focus on simplified versions of the algorithm, where the forest construction is independent of the training set. Consistency of such simplified models has been proved [e.g., Biau et al., 2008, Ishwaran and Kogalur, 2010, Denil et al., 2013]. However, these results do not extend to Breiman's [2001] original forests whose construction critically depends on the whole training set. Recent attempts to bridge the gap between theoretical forest models and Breiman's [2001] forests have been made by Wager [2014] and Scornet et al. [2015b] who establish consistency of the original algorithm under suitable assumptions.

Apart from the dependence of the forest construction on the data set, there is another fundamental difference between existing forest models and ones implemented. Indeed, in practice, a forest can only be grown with a finite number M of trees although most theoretical works assume, by convenience, that $M = \infty$. Since the predictor with $M = \infty$ does not depend on the specific tree realizations that form the forest, it is therefore more amenable to analysis. However, surprisingly, no study aims at clarifying the link between finite forests (finite M) and infinite forests ($M = \infty$) even if some authors [Mentch and Hooker, 2014a, Wager et al., 2014] proved results on finite forest predictions at a fixed point $\mathbf{x}$.

In the present paper, our goal is to study the connection between infinite forest models and finite forests used in practice in the context of regression. We start by proving a uniform central limit theorem for various random forests estimates, including Breiman's [2001] ones. In Section 3, assuming some regularity on the regression model, we point out that the $\mathbb{L}^2$ risk of any infinite forest is bounded above by the risk of the associated finite forests. Thus infinite forests are better estimate than finite forests in terms of $\mathbb{L}^2$ risk. Under the same assumptions, our analysis also shows that the risks of infinite and finite forests are close, if the number of trees is chosen to be large enough. An interesting corollary of this result is that infinite forest consistency implies finite forest consistency. Finally, in Section 4, we prove the consistency of several infinite random forests. In particular, taking one step toward the understanding of Breiman's [2001] forests, we prove that q quantile forests, a variety of forests whose construction depends on the positions $\mathbf{X}_i$'s of the data, are consistent. As for Breiman's [2001] forests, each leaf of each tree in q quantile forests contains a small number of points that does not grow to infinity with the sample size. Thus, q quantile forests average inconsistent trees estimate to build a consistent prediction.

We start by giving some notation in Section 2. All proofs are postponed to Section 5.

3.2 Notation

Throughout the paper, we assume to be given a training sample $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of $[0,1]^p \times \mathbb{R}$-valued independent random variables distributed as the prototype pair $(\mathbf{X}, Y)$, where $\mathbb{E}[Y^2] < \infty$. We aim at predicting the response Y, associated with the random variable $\mathbf{X}$, by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. In this context, we use random forests to build an estimate $m_n : [0,1]^p \to \mathbb{R}$ of m, based on the data set $\mathcal{D}_n$.

A random forest is a collection of M randomized regression trees [for an overview on tree construction, see Chapter 20 in Györfi et al., 2002]. For the j-th tree in the family, the predicted value at point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}, \Theta_j, \mathcal{D}_n)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ, independent of the sample $\mathcal{D}_n$. This random variable can be used to sample the training set or to select the candidate directions or positions for splitting. The trees are combined to form the finite forest estimate

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m). \tag{3.1}$$

By the law of large numbers, for any fixed $\mathbf{x}$, conditionally on $\mathcal{D}_n$, the finite forest estimate tends to the infinite forest estimate

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta [m_n(\mathbf{x}, \Theta)].$$

The risk of $m_{\infty,n}$ is defined by

$$R(m_{\infty,n}) = \mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2, \tag{3.2}$$

while the risk of $m_{M,n}$ equals

$$R(m_{M,n}) = \mathbb{E}[m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_M) - m(\mathbf{X})]^2. \tag{3.3}$$

It is stressed that both risks $R(m_{\infty,n})$ and $R(m_{M,n})$ are deterministic since the expectation in (3.2) is over $\mathbf{X}, \mathcal{D}_n$, and the expectation in (3.3) is over $\mathbf{X}, \mathcal{D}_n$ and $\Theta_1, \ldots, \Theta_M$. Throughout the paper, we say that $m_{\infty,n}$ (resp. $m_{M,n}$) is $\mathbb{L}^2$ consistent if $R(m_{\infty,n})$ (resp. $R(m_{M,n})$) tends to zero as $n \to \infty$.

As mentioned earlier, there is a large variety of forests, depending on how trees are grown and how the randomness Θ influences the tree construction. For instance, tree construction can be independent of $\mathcal{D}_n$ [Biau, 2012], depend only on the $\mathbf{X}_i$'s [Biau et al., 2008] or depend on the whole training set [Cutler and Zhao, 2001, Geurts et al., 2006, Zhu et al., 2012]. Throughout the paper, we use Breiman's [2001] forests and uniform forests to exemplify our results. In Breiman's [2001] original procedure, splits depend on the whole sample and are performed to minimize variance within the two resulting cells. The algorithm stops when each cell contains less than a small pre-specified number of points (typically, 5 in regression). On the other hand, uniform forests are a simpler procedure since, at each node, a coordinate is uniformly selected among $\{1, \ldots, p\}$ and a split position is uniformly chosen in the range of the cell, along the pre-chosen coordinate. The algorithm stops when a full binary tree of level k is built, that is if each cell has been cut exactly k times, where $k \in \mathbb{N}$ is a parameter of the algorithm.

In the rest of the paper, we will repeatedly use the random forest connection function K_n, defined as

$$\begin{array}{rccl} K_n: & [0,1]^p \times [0,1]^p & \to & [0,1] \\ & (\mathbf{x}, \mathbf{z}) & \mapsto & \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z} \right], \end{array}$$

where $\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z}$ is the event where $\mathbf{x}$ and $\mathbf{z}$ belong to the same cell in the tree $\mathcal{T}_n(\Theta)$ designed with Θ and $\mathcal{D}_n$. Moreover, notation $\mathbb{P}_\Theta$ denotes the probability with respect to Θ, conditionally on $\mathcal{D}_n$. The same notational convention holds for the expectation $\mathbb{E}_\Theta$ and the variance $\mathbb{V}_\Theta$. Thus, if we fix the training set $\mathcal{D}_n$, we see that the connection $K_n(\mathbf{x}, \mathbf{z})$ is just the probability that $\mathbf{x}$ and $\mathbf{z}$ are connected in the forest.

We say that a forest is discrete (resp. continuous) if, keeping $\mathcal{D}_n$ fixed, its connection function $K_n(\bullet, \bullet)$ is piecewise constant (resp. continuous). In fact, most existing forest models fall in one of these two categories. For example, if, at each cell, the number of possible splits is finite, then the forest is discrete. This is the case of Breiman's [2001] forests, where splits can only be performed at the middle of two consecutive data points along any coordinate. However, if splits are drawn according to some density along each coordinate, the resulting forest is continuous. For instance, uniform forests are continuous.

3.3 Finite and infinite random forests

Contrary to finite forests which depend upon the particular Θ_j's used to design trees, infinite forests do not and are therefore more amenable to mathematical analysis. Besides, finite forests predictions can be difficult to interpret since they depend on the random parameters Θ_j's. In addition, the Θ_j's are independent of the data set and thus unrelated to the particular prediction problem.

In this section, we study the link between finite forests and infinite forests. More specifically, assuming that the data set $\mathcal{D}_n$ is fixed, we examine the asymptotic behavior of the finite forest

estimate $m_{M,n}(\bullet, \Theta_1, \ldots, \Theta_M)$ as M tends to infinity. This setting is consistent with practical problems, where the $\mathcal{D}_n$ is fixed, and one can grow as many trees as possible.

Clearly, by the law of large numbers, we know that conditionally on $\mathcal{D}_n$, for all $\mathbf{x} \in [0,1]^p$, almost surely,

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) \underset{M\to\infty}{\to} m_{\infty,n}(\mathbf{x}). \tag{3.4}$$

The following theorem extend the pointwise convergence in (3.4) to the convergence of the whole functional estimate $m_{M,n}(\bullet, \Theta_1, \ldots, \Theta_M)$, towards the functional estimate $m_{\infty,n}(\bullet)$.

Theorem 3.1. *Consider a continuous or discrete random forest. Then, conditionally on $\mathcal{D}_n$, almost surely, for all $\mathbf{x} \in [0,1]^p$, we have*

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) \underset{M\to\infty}{\to} m_{\infty,n}(\mathbf{x}).$$

Remark 3.1. *Since the set $[0,1]^p$ is not countable, we cannot reverse the "almost sure" and "for all $\mathbf{x} \in [0,1]^p$" statements in (3.4). Thus, Theorem 3.1 is not a consequence of (3.4).*

Theorem 3.1 is a first step to prove that infinite forest estimates can be uniformly approximated by finite forest estimates. To pursue the analysis, a natural question is to determine the rate of convergence in Theorem 3.1. The pointwise rate of convergence is provided by the central limit theorem which says that, conditionally on $\mathcal{D}_n$, for all $\mathbf{x} \in [0,1]^p$,

$$\sqrt{M}(m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) - m_{\infty,n}(\mathbf{x})) \underset{M\to\infty}{\overset{\mathcal{D}}{\to}} \mathcal{N}(0, \tilde{\sigma}^2(\mathbf{x})), \tag{3.5}$$

where

$$\tilde{\sigma}^2(\mathbf{x}) = \mathbb{V}_\Theta\left(\frac{1}{N_n(\mathbf{x},\Theta)}\sum_{i=1}^n Y_i \mathbb{1}_{\mathbf{x}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\right) \leq 4 \max_{1\leq i\leq n} Y_i^2$$

(as before, $\mathbb{V}_\Theta$ denotes with respect to Θ, conditionally on $\mathcal{D}_n$), and $N_n(\mathbf{x}, \Theta)$ is the number of data points falling into the cell of the tree $\mathcal{T}_n(\Theta)$ which contains $\mathbf{x}$.

Equation (3.5) is not sufficient to determine the asymptotic distribution of the functional estimate $m_{M,n}(\bullet, \Theta_1, \ldots, \Theta_M)$. To make it explicit, we need to introduce the empirical process $\mathbb{G}_M$ [see van der Vaart and Wellner, 1996] defined by

$$\mathbb{G}_M = \sqrt{M}\left(\frac{1}{M}\sum_{m=1}^M \delta_{\Theta_m} - \mathbb{P}_\Theta\right),$$

where δ_{Θ_m} is the Dirac function at Θ_m. We also let $\mathcal{F}_2 = \{g_{\mathbf{x}} : \theta \mapsto m_n(\mathbf{x}, \theta); \mathbf{x} \in [0,1]^p\}$ be the collection of all possible tree estimates in the forest. In order to prove that a uniform central limit theorem holds for random forest estimates, we need to show that there exists a Gaussian process $\mathbb{G}$ such that

$$\sup_{g\in\mathcal{F}_2}\left\{\int_\Theta |g(\theta)|\mathrm{d}\mathbb{G}_M(\theta) - \int_\Theta |g(\theta)|\mathrm{d}\mathbb{G}(\theta)\right\} \underset{M\to\infty}{\to} 0, \tag{3.6}$$

where the first part on the left side can be written as

$$\int_\Theta |g(\theta)| \mathrm{d}\mathbb{G}_M(\theta) = \sqrt{M}\left(\frac{1}{M}\sum_{m=1}^{M}|g(\Theta_m)| - \mathbb{E}_\Theta\big[|g(\Theta)|\big]\right).$$

For more clarity, instead of (3.6), we will write

$$\sqrt{M}\left(\frac{1}{M}\sum_{m=1}^{M} m_n(\bullet,\Theta_m) - \mathbb{E}_\Theta\left[m_n(\bullet,\Theta)\right]\right) \xrightarrow{\mathcal{D}} \mathbb{G}g_\bullet. \tag{3.7}$$

To establish identity (3.7), we first define, for all $\varepsilon > 0$, the random forest grid step $\delta(\varepsilon)$ by

$$\delta(\varepsilon) = \sup\left\{\eta \in \mathbb{R} : \sup_{\substack{\mathbf{x}_1,\mathbf{x}_2\in[0,1]^p \\ \|\mathbf{x}_1-\mathbf{x}_2\|_\infty\leq\eta}} \big|1 - K_n(\mathbf{x}_1,\mathbf{x}_2)\big| \leq \frac{\varepsilon^2}{8}\right\},$$

where K_n is the connection function of the forest. The function δ can be seen as the modulus of continuity of K_n in the sense that it is the distance such that $K_n(\mathbf{x}_1,\mathbf{x}_2)$ does not vary of much that $\varepsilon^2/8$ if $\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \leq \delta(\varepsilon)$. We will also need the following assumption.

(H3.1) *One of the following properties is satisfied:*

- *The random forest is discrete,*
- *There exist $C, A > 0$, $\alpha < 2$ such that, for all $\varepsilon > 0$,*

$$\delta(\varepsilon) \geq C\exp(-A/\varepsilon^\alpha).$$

Observe that **(H3.1)** is mild since most forests are discrete and the only continuous forest we have in mind, the uniform forest, satisfies **(H3.1)**, as stated in Lemma 3.1 below.

Lemma 3.1. *Let $k \in \mathbb{N}$. Then, for all $\varepsilon > 0$, the grid step $\delta(\varepsilon)$ of uniform forests of level k satisfies*

$$\delta(\varepsilon) \geq \exp\left(-\frac{A_{k,p}}{\varepsilon^{2/3}}\right),$$

where $A_{k,p} = (8pe(k+2)!)^{1/3}$.

The following theorem states that a uniform central limit theorem is valid over the class of random forest estimates, providing that **(H3.1)** is satisfied.

Theorem 3.2. *Consider a random forest which satisfies* **(H3.1)**. *Then, conditionnally on $\mathcal{D}_n$,*

$$\sqrt{M}\left(m_{M,n}(\bullet) - m_{\infty,n}(\bullet)\right) \xrightarrow{\mathcal{D}} \mathbb{G}g_\bullet,$$

where $\mathbb{G}$ is a Gaussian process with mean zero and a covariate function

$$\mathrm{Cov}_\Theta(\mathbb{G}g_\mathbf{x}, \mathbb{G}g_\mathbf{z}) = \mathrm{Cov}_\Theta\left(\sum_{i=1}^{n} Y_i \frac{\mathbb{1}_{\mathbf{x}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{x},\Theta)}, \sum_{i=1}^{n} Y_i \frac{\mathbb{1}_{\mathbf{z}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{z},\Theta)}\right).$$

According to the discussion above, Theorem 3.2 holds for uniform forests (by Lemma 3.1) and Breiman's [2001] forests (since they are discrete). Moreover, according to this Theorem, the finite forest estimates tend uniformly to the infinite forest estimates, with the standard rate of convergence $\sqrt{M}$. This result contributes to bridge the gap between finite forests used in practice and infinite theoretical forests.

The proximity between two estimates can also be measured in terms of their $\mathbb{L}^2$ risk. In this respect, Theorem 3.3 states that the risk of infinite forests is lower than the one of finite forests and provides a bound on the difference between these two risks. We first need an assumption on the regression model.

(H3.2) *One has*

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a centered Gaussian noise with finite variance σ^2, independent of $\mathbf{X}$, and $\|m\|_\infty = \sup_{\mathbf{x}\in[0,1]^p} |m(\mathbf{x})| < \infty$.

Theorem 3.3. *Assume that* **(H3.2)** *is satisfied. Then, for all $M, n \in \mathbb{N}^\star$,*

$$R(m_{M,n}) = R(m_{\infty,n}) + \frac{1}{M}\,\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\Big[\mathbb{V}_\Theta\left[m_n(\mathbf{X},\Theta)\right]\Big].$$

In particular,

$$0 \leq R(m_{M,n}) - R(m_{\infty,n}) \leq \frac{8}{M} \times (\|m\|_\infty^2 + \sigma^2(1 + 4\log n)).$$

Theorem 3.3 reveals that the prediction accuracy of infinite forests is better than that of finite forests. In practice however, there is no simple way to implement infinite forests and, in fact, finite forests are nothing but Monte Carlo approximations of infinite forests. But, since the difference of risks between both types of forests is bounded (by Theorem 3.3), the prediction accuracy of finite forests is almost as good as that of infinite forests providing the number of trees is large enough. More precisely, under **(H3.2)**, for all $\varepsilon > 0$, if

$$M \geq \frac{8(\|m\|_\infty^2 + \sigma^2)}{\varepsilon} + \frac{32\sigma^2 \log n}{\varepsilon},$$

then $R(m_{M,n}) - R(m_{\infty,n}) \leq \varepsilon$.

Anoter interesting consequence of Theorem 3.3 is that, assuming that **(H3.2)** holds and that $M/\log n \to \infty$ as $n \to \infty$, finite random forests are consistent as soon as infinite random forests are. This alows to extend all previous consistency results regarding infinite forests [see, e.g., Meinshausen, 2006, Biau et al., 2008] to finite forests. It must be stressed that the "$\log n$" term comes from the Gaussian noise, since, if $\varepsilon_1, \ldots, \varepsilon_n$ are independent and distributed as a Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, we have,

$$\mathbb{E}\left[\max_{1\leq i\leq n} \varepsilon_i^2\right] \leq \sigma^2(1 + 4\log n),$$

[see, e.g., Chapter 1 in Boucheron et al., 2013]. Therefore, the required number of trees depends on the noise in the regression model. For instance, if Y is bounded, then the condition turns into $M \to \infty$.

3.4 Consistency of some random forest models

Section 3 was devoted to the connection between finite and infinite forests. In particular, we proved in Theorem 3.3 that the consistency of infinite forests implies that of finite forests, as soon as **(H3.2)** is satisfied and $M/\log n \to \infty$. Thus, it is natural to focus on the consistency of infinite forest estimates, which can be written as

$$m_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) Y_i, \tag{3.8}$$

where

$$W_{ni}^{\infty}(\mathbf{X}) = \mathbb{E}_{\Theta}\left[\frac{\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}}{N_n(\mathbf{X}, \Theta)}\right]$$

are the random forest weights.

3.4.1 Totally non adaptive forests

Proving consistency of infinite random forests is in general a difficult task, mainly because forest construction can depend on both the $\mathbf{X}_i$'s and the Y_i's. This feature makes the resulting estimate highly data-dependent, and therefore difficult to analyze (this is particularly the case for Breiman's [2001] forests). To simplify the analysis, we investigate hereafter infinite random forest estimates whose weights depends only on $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ which is called the X-property. The good news is that when infinite forest estimates have the X-property, they fall in the general class of local averaging estimates, whose consistency can be addressed using Stone's [1977] theorem.

Therefore, using Stone's theorem as a starting point, we first prove the consistency of random forests whose construction is independent of $\mathcal{D}_n$, which is the simplest case of random forests satisfying the X-property. For such forests, the construction is based on the random parameter Θ only. As for now, we say that a forest is totally non adaptive of level k ($k \in \mathbb{N}$, with k possibly depending on n) if each tree of the forest is built independently of the training set and if each cell is cut exactly k times. The resulting cell containing $\mathbf{X}$, designed with randomness Θ, is denoted by $A_n(\mathbf{X}, \Theta)$.

Theorem 3.4. *Assume that $\mathbf{X}$ is distributed on $[0,1]^p$ and consider a totally non adaptive forest of level k. In addition, assume that for all $\rho, \varepsilon > 0$, there exists $N > 0$ such that, with probability $1 - \rho$, for all $n > N$,*

$$diam(A_n(\mathbf{X}, \Theta)) \leq \varepsilon.$$

Then, providing $k \to \infty$ and $2^k/n \to 0$, the infinite random forest is $\mathbb{L}^2$ consistent, that is

$$R(m_{\infty,n}) \to 0 \quad as\ n \to \infty.$$

Theorem 3.4 is a generalization of some consistency results in Biau et al. [2008] for the case of totally non adaptive random forest. Together with Theorem 3.3, we see that if **(H3.2)** is satisfied and $M/\log n \to \infty$ as $n \to \infty$, then the finite random forest is $\mathbb{L}^2$ consistent.

According to Theorem 3.4, a totally non adaptive forest of level k is consistent if the cell diameters tend to zero as $n \to \infty$ and if the level k is properly tuned. This is in particular true for uniform random forests, as shown in the following corollary.

Corollary 3.1. *Assume that* $\mathbf{X}$ *is distributed on* $[0,1]^p$ *and consider a uniform forest of level* k*. Then, providing that* $k \to \infty$ *and* $2^k/n \to 0$*, the uniform random forest is* $\mathbb{L}^2$ *consistent.*

3.4.2 q quantile forests

For totally non adaptive forests, the main difficulty that consists in using the data set to build the forest and to predict at the same time, vanishes. However, because of their simplified construction, these forests are far from accurately modelling Breiman's [2001] forest. To take one step further into the understanding of Breiman's [2001] forest behavior, we study the q ($q \in [1/2, 1)$) quantile random forest, which satisfies the X-property. Indeed, their construction depends only on the X_i's which is a good trade off between the complexity of Breiman's [2001] forests and the simplicity of totally non adaptive forests. As an example of q quantile trees, the median tree ($q = 1/2$) has already been studied by Devroye et al. [1996], such as the k-spacing tree [Devroye et al., 1996] whose construction is based on quantiles.

Algorithm 3: q quantile forest predicted value at $\mathbf{x}$.

Input: Fix $a_n \in \{1, \dots, n\}$, and $\mathbf{x} \in [0,1]^p$.
Data: A training set $\mathcal{D}_n$.
for $\ell = 1, \dots, M$ **do**
  Select a_n points, without replacement, uniformly in $\mathcal{D}_n$.
  Set $\mathcal{P} = \{[0,1]^p\}$ the partition associated with the root of the tree.
  while *there exists* $A \in \mathcal{P}$ *which contains strictly more than two points* **do**
    Select uniformly one dimension j within $\{1, \dots, p\}$.
    Let N be the number of data points in A and select $q' \in [1-q, q] \cap (1/N, 1-1/N)$.
    Cut the cell A at the position given by the q' empirical quantile (see definition above) along the j-th coordinate.
    Call A_L and A_R the two resulting cell.
    Set $\mathcal{P} \leftarrow (\mathcal{P} \backslash \{A\}) \cup A_L \cup A_R$.
  end
  for *each* $A \in \mathcal{P}$ *which contains exactly two points* **do**
    Select uniformly one dimension j within $\{1, \dots, p\}$.
    Cut along the j-th direction, in the middle of the two points.
    Call A_L and A_R the two resulting cell.
    Set $\mathcal{P} \leftarrow (\mathcal{P} \backslash \{A\}) \cup A_L \cup A_R$.
  end
  Compute the predicted value $m_n(\mathbf{x}, \Theta_\ell)$ at $\mathbf{x}$ equal to the single Y_i falling in the cell of $\mathbf{x}$, with respect to the partition $\mathcal{P}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M)$ at the query point $\mathbf{x}$ according to equality (3.1).

In the spirit of Breiman's [2001] algorithm, before growing each tree, data are subsampled, that is a_n points ($a_n < n$) are selected without replacement. Then, each split is performed on an empirical q'-quantile (where $q' \in [1-q, q]$ can be pre-specified by the user or randomly chosen) along a coordinate, chosen uniformly at random among the p coordinates. Recall that the q'-quantile ($q' \in [1-q, q]$) of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is defined as the only $\mathbf{X}_{(\ell)}$ satisfying $F_n(\mathbf{X}_{(\ell-1)}) \leq q' < F_n(\mathbf{X}_{(\ell)})$, where the $\mathbf{X}_{(i)}$'s are ordered increasingly. Note that data points on which splits are performed are not sent down to the resulting cells. This is done to ensure that data points are uniformly distributed on the resulting cells (otherwise, there would be at least one data point on the edge of the resulting cell, and thus the data point distribution would not be uniform on this cell). Finally, the algorithm stops when each cell contains exactly one point. The full procedure is described in Algorithm 3.

Since the construction of q quantile forests depends on the $\mathbf{X}_i$'s and is based on subsampling, it is a more realistic modeling of Breiman's [2001] forests than totally non adaptive forests. It also provides a good understanding on why random forests are still consistent even when there is exactly one data point in each leaf. Theorem 3.5 states that with a proper subsampling rate of the training set, the q quantile random forests are consistent.

(H3.3) *One has*

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a centred noise such that $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \leq \sigma^2$, where $\sigma^2 < \infty$ is a constant. Moreover, $\mathbf{X}$ has a density on $[0,1]^p$ and m is continuous.

Theorem 3.5. *Assume that* **(H3.3)** *is satisfied. Then, providing $a_n \to \infty$ et $a_n/n \to 0$, the infinite q quantile random forest is $\mathbb{L}^2$ consistent.*

3.4.3 Discussion

Some remarks are in order. At first, observe that each tree in the q quantile forest is inconsistent [see Problem 4.3 in Györfi et al., 2002], because each leaf contains exactly one data point, a number which does not grow to infinity as $n \to \infty$. Thus, Theorem 3.5 shows that q quantile forest combines inconsistent trees to form a consistent estimate.

Secondly, many random forests can be seen as quantile forests if they satisfy the X-property and if splits do not separate a small fraction of data points from the rest of the sample (indeed, for each split in the q quantile forests, the resulting cells contain at least a fraction q of the observations falling into the parent node). The last assumption is true, for example, if $\mathbf{X}$ has a density on $[0,1]^p$ bounded from below and from above, and if some splitting rule forces splits to be performed far away from the cell edges. This assumption is explicitly made in the analysis of Meinshausen [2006] and Wager [2014] to ensure that cell diameters tend to zero as $n \to \infty$, which is a necessary condition to prove the consistency of partitioning estimates [see Chapter 4 in Györfi et al., 2002]. Unfortunately, there are no results stating that splits in Breiman's [2001] forests are performed far from the edges [see Ishwaran, 2013, for an analysis of the splitting criterion in Breiman's forests].

In addition, we note that Theorem 3.5 does not cover the bootstrap case since in that case, $a_n = n$ data points are selected with replacement. However, the condition on the subsampling

rate can be replaced by the following one: for all $\mathbf{x}$,

$$\max_i \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i \right] \to 0 \text{ as } n \to \infty. \tag{3.9}$$

Condition (3.9) can be interpreted by saying that a point $\mathbf{x}$ should not be connected too often to the same data point in the forest, thus meaning that trees have to be various enough to ensure the forest consistency. This idea of diversity among trees has already been suggested by Breiman [2001]. In bootstrap case, a single data point is selected in about 63% of trees. Thus, the term $\max_i \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i \right]$ is roughly upper bounded by 0.63 which is not sufficient to prove (3.9). It does not mean that random forests based on bootstrap are inconsistent but that a more detailed analysis is required. A possible, but probably difficult, route is an in-depth analysis of the connection function $K_n(\mathbf{x}, \mathbf{X}_i) = \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i \right]$.

Finally, a natural question is how to extend random forests to the case of functional data [see, e.g., Ramsay and Silverman, 2005, Ferraty and Vieu, 2006, Horváth and Kokoszka, 2012, Bongiorno et al., 2014, for an overview of functional data analysis]. A first attempt may be done by expanding each variable in a particular truncated functional basis. Each curve is then represented by a finite number of coefficient and any standard random forest procedure can be applied [see, e.g., Poggi and Tuleau, 2006, Gregorutti et al., 2014, for practical applications]. Since this method mainly consists in projecting functional variables onto finite dimensional spaces, it suffers from several drawbacks (for example, it depends on the basis and on the truncated procedure which are arbitrarily chosen in most cases). Unfortunately, we are not aware of functional random forest procedures that can directly handle functional data. Given the good performance of random forests in high dimensional settings and the numerous applications involving functional data, developing such functional forests is certainly is an interesting research topic.

3.5 Proofs

3.5.1 Proof of Theorem 3.1

Note that the forest estimate $m_{M,n}$ can be written as

$$m_{M,n}(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}^{M}(\mathbf{x}) Y_i,$$

where

$$W_{ni}^{M}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{X}_i}}{N_n(\mathbf{x}, \Theta_m)}.$$

Similarly, one can write the infinite forest estimate $m_{\infty,n}$ as

$$m_{\infty,n}(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{x}) Y_i,$$

where

$$W_{ni}^{\infty}(\mathbf{x}) = \mathbb{E}_{\Theta}\left[\frac{\mathbb{1}_{\mathbf{x}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{x},\Theta)}\right].$$

We assume that $\mathcal{D}_n$ is fixed and prove Theorem 3.1 for $p = 2$. The general case can be treated similarly. Throughout the proof, we write, for all θ, $\mathbf{x}, \mathbf{z} \in [0,1]^2$,

$$f_{\mathbf{x},\mathbf{z}}(\theta) = \frac{\mathbb{1}_{\mathbf{x}\overset{\theta}{\leftrightarrow}\mathbf{z}}}{N_n(\mathbf{x},\theta)}.$$

Let us first consider a discrete random forest. By definition of such random forests, there exists $v \in \mathbb{N}^{\star}$ and a partition $\{A_\ell : 1 \le \ell \le v\}$ of $[0,1]^2$ such that the connection function K_n is constant over the sets $A_{\ell_1} \times A_{\ell_2}$'s $(1 \le \ell_1, \ell_2 \le v)$. For all $1 \le \ell \le v$, denote by a_ℓ, the center of the cell A_ℓ. Take $\mathbf{x}, \mathbf{z} \in [0,1]^2$. There exist ℓ_1, ℓ_2 such that $\mathbf{x} \in A_{\ell_1}, \mathbf{z} \in A_{\ell_2}$. Thus, for all θ,

$$\begin{aligned}
\left|\frac{\mathbb{1}_{\mathbf{x}\overset{\theta}{\leftrightarrow}\mathbf{z}}}{N_n(\mathbf{x},\theta)} - \frac{\mathbb{1}_{\mathbf{a}_{\ell_1}\overset{\theta}{\leftrightarrow}\mathbf{a}_{\ell_2}}}{N_n(\mathbf{a}_{\ell_1},\theta)}\right| &\le \left|\frac{\mathbb{1}_{\mathbf{x}\overset{\theta}{\leftrightarrow}\mathbf{z}}}{N_n(\mathbf{x},\theta)} - \frac{\mathbb{1}_{\mathbf{a}_{\ell_1}\overset{\theta}{\leftrightarrow}\mathbf{z}}}{N_n(\mathbf{a}_{\ell_1},\theta)} + \frac{\mathbb{1}_{\mathbf{a}_{\ell_1}\overset{\theta}{\leftrightarrow}\mathbf{z}}}{N_n(\mathbf{a}_{\ell_1},\theta)} - \frac{\mathbb{1}_{\mathbf{a}_{\ell_1}\overset{\theta}{\leftrightarrow}\mathbf{a}_{\ell_2}}}{N_n(\mathbf{a}_{\ell_1},\theta)}\right| \\
&\le \frac{1}{N_n(\mathbf{a}_{\ell_1},\theta)}\left|\mathbb{1}_{\mathbf{x}\overset{\theta}{\leftrightarrow}\mathbf{z}} - \mathbb{1}_{\mathbf{a}_{\ell_1}\overset{\theta}{\leftrightarrow}\mathbf{z}}\right| \\
&\quad + \frac{1}{N_n(\mathbf{a}_{\ell_1},\theta)}\left|\mathbb{1}_{\mathbf{a}_{\ell_1}\overset{\theta}{\leftrightarrow}\mathbf{z}} - \mathbb{1}_{\mathbf{a}_{\ell_1}\overset{\theta}{\leftrightarrow}\mathbf{a}_{\ell_2}}\right| \\
&\le \frac{1}{N_n(\mathbf{a}_{\ell_1},\theta)}\mathbb{1}_{\mathbf{x}\overset{\theta}{\leftrightarrow}\mathbf{a}_{\ell_1}} + \frac{1}{N_n(\mathbf{a}_{\ell_1},\theta)}\mathbb{1}_{\mathbf{a}_{\ell_2}\overset{\theta}{\leftrightarrow}\mathbf{z}} \\
&\le 0.
\end{aligned}$$

Thus, the set

$$\mathcal{H} = \left\{\theta \mapsto f_{\mathbf{x},\mathbf{z}}(\theta) : \mathbf{x}, \mathbf{z} \in [0,1]^2\right\}$$

is finite. Therefore, by the strong law of large numbers, almost surely, for all $f \in \mathcal{H}$,

$$\frac{1}{M}\sum_{m=1}^{M} f(\Theta_m) \underset{M\to\infty}{\to} \mathbb{E}_{\Theta}\big[f(\Theta)\big].$$

Noticing that $W_{ni}^{M}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} f_{\mathbf{x},\mathbf{X}_i}(\Theta_m)$, we obtain that, almost surely, for all $\mathbf{x} \in [0,1]^2$,

$$W_{ni}^{M}(\mathbf{x}) \to W_{ni}^{\infty}(\mathbf{x}), \quad \text{as } M \to \infty.$$

Since $\mathcal{D}_n$ is fixed and random forest estimates are linear in the weights, the proof of the discrete case is complete.

Let us now consider a continuous random forest. We define, for all $\mathbf{x}, \mathbf{z} \in [0,1]^2$,

$$W_{n}^{M}(\mathbf{x},\mathbf{z}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\mathbb{1}_{\mathbf{x}\overset{\Theta_m}{\leftrightarrow}\mathbf{z}}}{N_n(\mathbf{x},\Theta_m)},$$

and

$$W_n^{\infty}(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\Theta}\left[\frac{\mathbb{1}_{\mathbf{x}\overset{\Theta}{\leftrightarrow}\mathbf{z}}}{N_n(\mathbf{x},\Theta)}\right].$$

According to the strong law of large numbers, almost surely, for all $\mathbf{x},\mathbf{z}\in[0,1]^2\cap\mathbb{Q}^2$,

$$\lim_{M\to\infty} W_n^M(\mathbf{x},\mathbf{z}) = W_n^{\infty}(\mathbf{x},\mathbf{z}).$$

Set $\mathbf{x},\mathbf{z}\in[0,1]^2$ where $\mathbf{x}=(x^{(1)},x^{(2)})$ and $\mathbf{z}=(z^{(1)},z^{(2)})$. Assume, without loss of generality, that $x^{(1)}<z^{(1)}$ and $x^{(2)}<z^{(2)}$. Let

$$A_{\mathbf{x}} = \{\mathbf{u}\in[0,1]^2, u^{(1)}\le x^{(1)} \text{ and } u^{(2)}\le x^{(2)}\},$$
$$\text{and } A_{\mathbf{z}} = \{\mathbf{u}\in[0,1]^2, u^{(1)}\ge z^{(1)} \text{ and } u^{(2)}\ge z^{(2)}\}.$$

Choose $\mathbf{x}_1\in A_{\mathbf{x}}\cap\mathbb{Q}^2$ (resp. $\mathbf{z}_2\in A_{\mathbf{z}}\cap\mathbb{Q}^2$) and take $\mathbf{x}_2\in[0,1]^2\cap\mathbb{Q}^2$ (resp. $\mathbf{z}_1\in[0,1]^2\cap\mathbb{Q}^2$) such that $\mathbf{x}_1,\mathbf{x},\mathbf{x}_2$ (resp. $\mathbf{z}_2,\mathbf{z},\mathbf{z}_1$) are aligned in this order (see Figure 3.1).

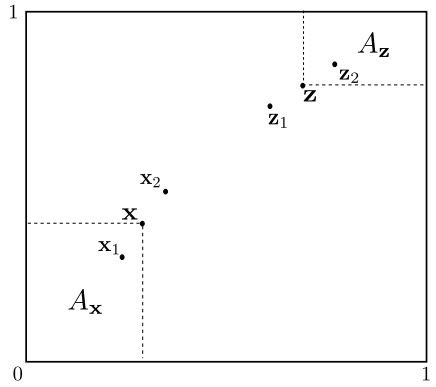

Figure 3.1: Respective positions of $\mathbf{x},\mathbf{x}_1,\mathbf{x}_2$ and $\mathbf{z},\mathbf{z}_1,\mathbf{z}_2$

Thus,

$$\begin{aligned}\left|W_n^M(\mathbf{x},\mathbf{z}) - W_n^{\infty}(\mathbf{x},\mathbf{z})\right| \le &\left|W_n^M(\mathbf{x},\mathbf{z}) - W_n^M(\mathbf{x}_1,\mathbf{z}_2)\right| \\ &+ \left|W_n^M(\mathbf{x}_1,\mathbf{z}_2) - W_n^{\infty}(\mathbf{x}_1,\mathbf{z}_2)\right| \\ &+ \left|W_n^{\infty}(\mathbf{x}_1,\mathbf{z}_2) - W_n^{\infty}(\mathbf{x},\mathbf{z})\right|. \end{aligned} \tag{3.10}$$

Set $\varepsilon>0$. Because of the continuity of K_n, we can choose $\mathbf{x}_1,\mathbf{x}_2$ close enough to $\mathbf{x}$ and $\mathbf{z}_2,\mathbf{z}_1$ close enough to $\mathbf{z}$ such that,

$$\begin{aligned}|K_n(\mathbf{x}_2,\mathbf{x}_1)-1| &\le \varepsilon,\\ |K_n(\mathbf{z}_1,\mathbf{z}_2)-1| &\le \varepsilon,\\ |1-K_n(\mathbf{x}_1,\mathbf{x})| &\le \varepsilon,\\ |1-K_n(\mathbf{z}_2,\mathbf{z})| &\le \varepsilon.\end{aligned}$$

Let us consider the second term in equation (3.10). Since $\mathbf{x}_1, \mathbf{z}_2$ belong to $[0,1]^2 \cap \mathbb{Q}^2$, almost surely, there exists $M_1 > 0$ such that, if $M > M_1$,

$$\left|W_n^M(\mathbf{x}_1, \mathbf{z}_2) - W_n^\infty(\mathbf{x}_1, \mathbf{z}_2)\right| \leq \varepsilon.$$

Regarding the first term in (3.10), note that, according to the position of $\mathbf{x}_1, \mathbf{z}_2, \mathbf{x}$, for all θ, we have

$$\frac{\mathbb{1}_{\mathbf{x}_1 \overset{\theta}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}, \theta)} = \frac{\mathbb{1}_{\mathbf{x}_1 \overset{\theta}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}_1, \theta)}.$$

Therefore

$$\left|W_n^M(\mathbf{x}, \mathbf{z}) - W_n^M(\mathbf{x}_1, \mathbf{z}_2)\right| \leq \frac{1}{M} \sum_{m=1}^{M} \left| \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}}}{N_n(\mathbf{x}, \Theta_m)} - \frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}, \Theta_m)} \right|.$$

Observe that, given the positions of $\mathbf{x}, \mathbf{x}_1, \mathbf{z}, \mathbf{z}_2$, the only case where

$$\left| \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}}}{N_n(\mathbf{x}, \Theta_m)} - \frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}, \Theta_m)} \right| \neq 0$$

occurs when $\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_2$ and $\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}$. Thus,

$$\begin{aligned}
&\frac{1}{M} \sum_{m=1}^{M} \left| \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}}}{N_n(\mathbf{x}, \Theta_m)} - \frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}, \Theta_m)} \right| \\
&\quad = \frac{1}{M} \sum_{m=1}^{M} \left| \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}}}{N_n(\mathbf{x}, \Theta_m)} - \frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}, \Theta_m)} \right| \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_2} \mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}} \\
&\quad \leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}} \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_2}.
\end{aligned}$$

Again, given the relative positions of $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}, \mathbf{z}_2, \mathbf{z}_1$, we obtain

$$\begin{aligned}
\frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}} \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_2} &\leq \frac{1}{M} \sum_{m=1}^{M} \left(\mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{x}} + \mathbb{1}_{\mathbf{z}_2 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}} \right) \\
&\leq \frac{1}{M} \sum_{m=1}^{M} \left(\mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{x}_2} + \mathbb{1}_{\mathbf{z}_2 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_1} \right) \\
&\leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{x}_2} + \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{z}_2 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_1}.
\end{aligned}$$

Collecting the previous inequalities, we have

$$\begin{aligned}
\left|W_n^M(\mathbf{x}, \mathbf{z}) - W_n^\infty(\mathbf{x}_1, \mathbf{z}_2)\right| &\leq \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\nleftrightarrow} \mathbf{x}_2} + \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{z}_2 \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}_1} \\
&\leq 2 - \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{x}_1 \overset{\Theta_m}{\leftrightarrow} \mathbf{x}_2} - \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\mathbf{z}_2 \overset{\Theta_m}{\leftrightarrow} \mathbf{z}_1}.
\end{aligned}$$

Since $\mathbf{x}_2, \mathbf{z}_1, \mathbf{x}_1, \mathbf{z}_2 \in [0,1]^2 \cap \mathbb{Q}^2$, we deduce that there exists M_2 such that, for all $M > M_2$,

$$\left|W_n^M(\mathbf{x}, \mathbf{z}) - W_n^\infty(\mathbf{x}_1, \mathbf{z}_2)\right| \leq 2 - K_\infty(\mathbf{x}_2, \mathbf{x}_1) - K_\infty(\mathbf{z}_1, \mathbf{z}_2) + 2\varepsilon. \tag{3.11}$$

Considering the third term in (3.10), using the same arguments as above, we see that

$$\begin{aligned}
|W_n^\infty(\mathbf{x}_1, \mathbf{z}_2) - W_n^\infty(\mathbf{x}, \mathbf{z})| &\leq \mathbb{E}_\Theta \left| \frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}_1, \Theta)} - \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z}}}{N_n(\mathbf{x}, \Theta)} \right| \\
&\leq \mathbb{E}_\Theta \left[\left| \frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}_1, \Theta)} - \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z}}}{N_n(\mathbf{x}, \Theta)} \right| \mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\leftrightarrow} \mathbf{z}_2} \mathbb{1}_{\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z}} \right] \\
&\leq \mathbb{E}_\Theta \left[\mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\nleftrightarrow} \mathbf{z}_2} \mathbb{1}_{\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z}} \right] \\
&\leq \mathbb{E}_\Theta \left[\mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\nleftrightarrow} \mathbf{x}_2} + \mathbb{1}_{\mathbf{z}_2 \overset{\Theta}{\nleftrightarrow} \mathbf{z}_1} \right] \\
&\leq 2 - K_n(\mathbf{x}_1, \mathbf{x}_2) - K_n(\mathbf{z}_2, \mathbf{z}_1).
\end{aligned} \tag{3.12}$$

Using inequalities (3.11) and (3.12) in (3.10), we finally conclude that, for all $M > \max(M_1, M_2)$,

$$\begin{aligned}
\left|W_n^M(\mathbf{x}, \mathbf{z}) - W_n^\infty(\mathbf{x}, \mathbf{z})\right| &\leq 4 - 2K_n(\mathbf{x}_2, \mathbf{x}_1) - 2K_n(\mathbf{z}_1, \mathbf{z}_2) + 3\varepsilon \\
&\leq 7\varepsilon.
\end{aligned}$$

This completes the proof of Theorem 3.1.

3.5.2 Proof of Lemma 3.1 and Theorem 3.2

Proof of Lemma 3.1. Set $k \in \mathbb{N}$ and $\varepsilon > 0$. We start by considering the case where $p = 1$. Take $x, z \in [0,1]$ and let $w = -\log(|x - z|)$. The probability that x and z are not connected in the uniform forest after k cuts is given by

$$\begin{aligned}
1 - K_k(x, z) &\leq 1 - K_k(0, |z - x|) \\
&\qquad \text{(according to Technical Lemma 3.1, see the end of the section)} \\
&\leq e^{-w} \mathbb{1}_{k>0} \sum_{i=0}^{k-1} \frac{w^i}{i!} \\
&\qquad \text{(according to Technical Lemma 3.2, see the end of the section)} \\
&\leq \frac{(k+2)!e}{w^3},
\end{aligned}$$

for all $w > 1$. Now, consider the multivariate case, and let $\mathbf{x}, \mathbf{z} \in [0,1]^p$. Set, for all $1 \leq j \leq p$, $w_j = -\log(|x_j - z_j|)$. By union bound, recalling that $1 - K_k(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta(\mathbf{x} \overset{\Theta}{\nleftrightarrow} \mathbf{z})$, we have

$$\begin{aligned}
1 - K_k(\mathbf{x}, \mathbf{z}) &\leq \sum_{j=1}^{p} (1 - K_k(x_j, z_j)) \\
&\leq \frac{p(k+2)!e}{\min_{1 \leq j \leq p} w_j^3}.
\end{aligned}$$

Thus, if, for all $1 \leq j \leq p$,

$$|x_j - z_j| \leq \exp\left(-\frac{(A_{k,p})^{1/3}}{\varepsilon^{2/3}}\right),$$

then

$$1 - K_k(\mathbf{x}, \mathbf{z}) \leq \frac{\varepsilon^2}{8},$$

where $A_{k,p} = (8pe(k+2)!)^{1/3}$. Consequently,

$$\delta(\varepsilon) \geq \exp\left(-\frac{(A_{k,p})^{1/3}}{\varepsilon^{2/3}}\right).$$

□

Proof of Theorem 3.2. We start the proof by proving that the class

$$\mathcal{H} = \left\{\theta \mapsto f_{\mathbf{x},\mathbf{z}}(\theta) : \mathbf{x}, \mathbf{z} \in [0,1]^2\right\}$$

is $\mathbb{P}_\Theta$-Donsker, that is, there exists a Gaussian process $\mathbb{G}$ such that

$$\sup_{f \in \mathcal{H}} \left\{\mathbb{E}|f| \left(d\mathbb{G}_M - d\mathbb{G}\right)\right\} \underset{M \to \infty}{\to} 0.$$

At first, let us consider a finite random forest. As noticed in the proof of Theorem 3.1, the set $\mathcal{H}$ is finite. Consequently, by the central limit theorem, the set $\mathcal{H}$ is $\mathbb{P}_\Theta$-Donsker.

Now, consider a random forest which satisfies the second statement in **(H3.1)**. Set $\varepsilon > 0$. Consider a regular grid of $[0,1]^p$ with a step δ and let $\mathcal{G}_\delta$ be the set of nodes of this grid. We start by finding a condition on δ such that the set

$$\tilde{\mathcal{G}}_\delta = \{[f_{\mathbf{x}_1,\mathbf{z}_1}, f_{\mathbf{x}_2,\mathbf{z}_2}] : \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{G}_\delta\}$$

is a covering of ε-bracket of the set $\mathcal{H}$, that is, for all $f \in \mathcal{H}$, there exists $\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2 \in \mathcal{G}_\delta$ such that

$$f_{\mathbf{x}_1,\mathbf{z}_1} \leq f \leq f_{\mathbf{x}_2,\mathbf{z}_2} \text{ and } \mathbb{E}^{1/2}\left[f_{\mathbf{x}_2,\mathbf{z}_2}(\Theta) - f_{\mathbf{x}_1,\mathbf{z}_1}(\Theta)\right]^2 \leq \varepsilon. \tag{3.13}$$

To this aim, set $\mathbf{x}, \mathbf{z} \in [0,1]^p$ and choose $\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{G}_\delta$ (see Figure 3.2). Note that, for all θ,

$$\frac{\mathbb{1}_{\mathbf{x}_1 \overset{\theta}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}_1, \theta)} \leq \frac{\mathbb{1}_{\mathbf{x} \overset{\theta}{\leftrightarrow} \mathbf{z}}}{N_n(\mathbf{x}, \theta)} \leq \frac{\mathbb{1}_{\mathbf{x}_2 \overset{\theta}{\leftrightarrow} \mathbf{z}_1}}{N_n(\mathbf{x}_2, \theta)},$$

that is, $f_{\mathbf{x}_1,\mathbf{z}_2} \leq f_{\mathbf{x},\mathbf{z}} \leq f_{\mathbf{x}_2,\mathbf{z}_1}$. To prove the second statement in (3.13), observe that

$$\begin{aligned}
\mathbb{E}^{1/2}\big[f_{\mathbf{x}_2,\mathbf{z}_2}(\Theta) - f_{\mathbf{x}_1,\mathbf{z}_1}(\Theta)\big]^2 &= \mathbb{E}_\Theta^{1/2}\left[\frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}_1,\Theta)} - \frac{\mathbb{1}_{\mathbf{x}_2 \overset{\Theta}{\leftrightarrow} \mathbf{z}_1}}{N_n(\mathbf{x}_2,\Theta)}\right]^2 \\
&= \mathbb{E}_\Theta^{1/2}\Bigg[\left(\frac{\mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\leftrightarrow} \mathbf{z}_2}}{N_n(\mathbf{x}_1,\Theta)} - \frac{\mathbb{1}_{\mathbf{x}_2 \overset{\Theta}{\leftrightarrow} \mathbf{z}_1}}{N_n(\mathbf{x}_2,\Theta)}\right) \\
&\qquad \times \mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\nleftrightarrow} \mathbf{z}_2} \mathbb{1}_{\mathbf{x}_2 \overset{\Theta}{\leftrightarrow} \mathbf{z}_1}\Bigg]^2 \\
&\leq \mathbb{E}_\Theta^{1/2}\left[\mathbb{1}_{\mathbf{x}_1 \overset{\Theta}{\nleftrightarrow} \mathbf{x}_2} + \mathbb{1}_{\mathbf{z}_1 \overset{\Theta}{\nleftrightarrow} \mathbf{z}_2}\right]^2 \\
&\leq 2\sqrt{1 - K_n(\mathbf{x}_1,\mathbf{x}_2) + 1 - K_n(\mathbf{z}_1,\mathbf{z}_2)}.
\end{aligned}$$

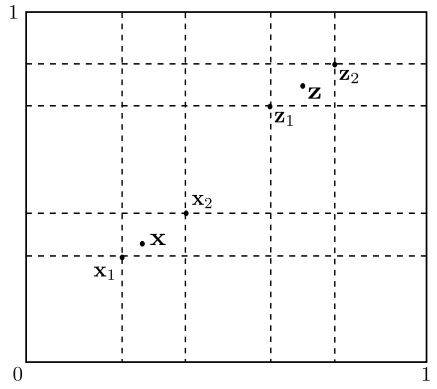

Figure 3.2: Respective positions of $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2$ with $p = 2$.

Thus, we have to choose the grid step δ such that

$$\sup_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in [0,1]^p \\ \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \leq \delta}} |1 - K_n(\mathbf{x}_1,\mathbf{x}_2)| \leq \frac{\varepsilon^2}{8}. \tag{3.14}$$

By **(H3.1)** and the definition of the random forest grid step, there exist constants $C, A > 0$ and $0 < \alpha < 2$ such that, for all $\varepsilon > 0$, if

$$\delta \geq C \exp(-A/\varepsilon^\alpha), \tag{3.15}$$

then (3.14) is satisfied. Hence, if δ satisfies (3.15), then $\tilde{\mathcal{G}}_\delta$ is a covering of ε-bracket of $\mathcal{H}$. In that case, the number $N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P))$ of ε-bracket needed to cover $\mathcal{H}$ satisfies

$$N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P)) \leq \mathrm{Card}(\tilde{\mathcal{G}}_\delta) \leq \mathrm{Card}(\mathcal{G}_\delta)^4 \leq \left(\frac{1}{\delta}\right)^{4p}.$$

Consequently,

$$\sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P))} \leq \sqrt{\frac{2Ap}{\varepsilon^\alpha} - 2p \log C}$$

where the last term is integrable near zero since $\alpha < 2$. Thus, according to Theorem 2.5.6 in van der Vaart and Wellner [1996] (and the remark at the beginning of Section 2.5.2), the class $\mathcal{H}$ is $\mathbb{P}_\Theta$-Donsker.

To conclude the proof, consider a random forest satisfying **(H3.1)**. From above, we see that the class $\mathcal{H}$ is $\mathbb{P}_\Theta$-Donsker. Recall that $\mathcal{F}_2 = \{g_\mathbf{x} : \theta \mapsto m_n(\mathbf{x}, \theta) : \mathbf{x} \in [0,1]^p\}$, where

$$m_n(\mathbf{x}, \Theta) = \sum_{i=1}^n Y_i f_{\mathbf{x},\mathbf{X}_i}(\Theta).$$

Since the training set $\mathcal{D}_n$ is fixed, we have

$$\begin{aligned}
&\sup_{g_\mathbf{x} \in \mathcal{F}_2} \left\{ \mathbb{E}|g_\mathbf{x}| \, (d\mathbb{G}_M - d\mathbb{G}) \right\} \\
&\quad = \sup_{\mathbf{x} \in [0,1]^p} \left\{ \mathbb{E}\Big| \sum_{i=1}^n Y_i f_{\mathbf{x},\mathbf{X}_i} \Big| \, (d\mathbb{G}_M - d\mathbb{G}) \right\} \\
&\quad \leq \sum_{i=1}^n |Y_i| \sup_{\mathbf{x} \in [0,1]^p} \left\{ \mathbb{E}|f_{\mathbf{x},\mathbf{X}_i}| \, (d\mathbb{G}_M - d\mathbb{G}) \right\} \\
&\quad \leq \left(\sum_{i=1}^n |Y_i| \right) \sup_{\mathbf{x},\mathbf{z} \in [0,1]^p} \Big\{ \mathbb{E}|f_{\mathbf{x},\mathbf{z}}| \, (d\mathbb{G}_M - d\mathbb{G}) \Big\},
\end{aligned}$$

which tends to zero as M tends to infinity, since the class $\mathcal{H}$ is $\mathbb{P}_\Theta$-Donsker.

Finally, note that Breiman's [2001] random forests are discrete, thus satisfying **(H3.1)**. Uniform forests are continuous and satisfy **(H3.1)** according to Lemma 3.1.

□

3.5.3 Proof of Theorem 3.3

Observe that,

$$\begin{aligned}
&\Big(m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_m) - m(\mathbf{X}) \Big)^2 \\
&\quad = \Big(m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_m) - \mathbb{E}_\Theta \left[m_n(\mathbf{X}, \Theta) \right] \Big)^2 + \Big(\mathbb{E}_\Theta \left[m_n(\mathbf{X}, \Theta) \right] - m(\mathbf{X}) \Big)^2 \\
&\qquad + 2 \Big(\mathbb{E}_\Theta \left[m_n(\mathbf{X}, \Theta) \right] - m(\mathbf{X}) \Big) \Big(m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_m) - \mathbb{E}_\Theta \left[m_n(\mathbf{X}, \Theta) \right] \Big).
\end{aligned}$$

Taking the expectation on both sides, we obtain

$$R(m_{M,n}) = R(m_{\infty,n}) + \mathbb{E}\Big[m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_m) - \mathbb{E}_\Theta \left[m_n(\mathbf{X}, \Theta) \right] \Big]^2,$$

by noticing that

$$\begin{aligned}
&\mathbb{E}\Big[\Big(m_{M,n}(\mathbf{X},\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\left[m_n(\mathbf{X},\Theta)\right]\Big)\Big(\mathbb{E}_\Theta\left[m_n(\mathbf{X},\Theta)\right] - m(\mathbf{X})\Big)\Big] \\
&\quad = \mathbb{E}_{\mathbf{X},\mathcal{D}_n}\Bigg[\Big(\mathbb{E}_\Theta\left[m_n(\mathbf{X},\Theta)\right] - m(\mathbf{X})\Big) \\
&\qquad\qquad \times \mathbb{E}_{\Theta_1,\ldots,\Theta_M}\Big[m_{M,n}(\mathbf{X},\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\big[m_n(\mathbf{X},\Theta)\big]\Big]\Bigg] \\
&\quad = 0,
\end{aligned}$$

according to the definition of $m_{M,n}$. Fixing $\mathbf{X}$ and $\mathcal{D}_n$, note that random variables $m_n(\mathbf{X},\Theta_1)$, $\ldots, m_n(\mathbf{X},\Theta_1)$ are independent and identically distributed. Thus, we have

$$\begin{aligned}
&\mathbb{E}\left[m_{M,n}(\mathbf{X},\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\left[m_n(\mathbf{X},\Theta)\right]\right]^2 \\
&= \mathbb{E}_{\mathbf{X},\mathcal{D}_n}\mathbb{E}_{\Theta_1,\ldots,\Theta_M}\left[\frac{1}{M}\sum_{m=1}^{M} m_n(\mathbf{X},\Theta_m) - \mathbb{E}_\Theta\left[m_n(\mathbf{X},\Theta)\right]\right]^2 \\
&= \frac{1}{M} \times \mathbb{E}\Big[\mathbb{V}_\Theta\left[m_n\left(\mathbf{X},\Theta\right)\right]\Big],
\end{aligned}$$

which conludes the first part of the proof. Now, note that the tree estimate $m_n(\mathbf{X},\Theta)$ can be written as

$$m_n(\mathbf{X},\Theta) = \sum_{i=1}^{n} W_{ni}(\mathbf{X},\Theta) Y_i,$$

where

$$W_{ni}(\mathbf{X},\Theta) = \frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}.$$

Therefore,

$$\begin{aligned}
R(m_{M,n}) - R(m_{\infty,n}) &= \frac{1}{M} \times \mathbb{E}\Big[\mathbb{V}_\Theta\left[m_n(\mathbf{X},\Theta)\right]\Big] \\
&= \frac{1}{M} \times \mathbb{E}\left[\mathbb{V}_\Theta\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X},\Theta)(m(\mathbf{X}_i)+\varepsilon_i)\right]\right] \\
&\leq \frac{1}{M} \times \mathbb{E}\left[\mathbb{E}_\Theta\left[\max_{1\leq i\leq n}(m(\mathbf{X}_i)+\varepsilon_i) - \min_{1\leq j\leq n}(m(\mathbf{X}_j)+\varepsilon_j)\right]^2\right] \\
&\leq \frac{1}{M} \times \mathbb{E}\Bigg[2\mathbb{E}_\Theta\left[\max_{1\leq i\leq n} m(\mathbf{X}_i) - \min_{1\leq j\leq n} m(\mathbf{X}_j)\right]^2 \\
&\qquad + 2\mathbb{E}_\Theta\left[\max_{1\leq i\leq n}\varepsilon_i - \min_{1\leq j\leq n}\varepsilon_j\right]^2\Bigg] \\
&\leq \frac{1}{M} \times \left[8\|m\|_\infty^2 + 2\mathbb{E}\left[\max_{1\leq i\leq n}\varepsilon_i - \min_{1\leq j\leq n}\varepsilon_j\right]^2\right] \\
&\leq \frac{1}{M} \times \left[8\|m\|_\infty^2 + 8\sigma^2\mathbb{E}\left[\max_{1\leq i\leq n}\frac{\varepsilon_i}{\sigma}\right]^2\right].
\end{aligned}$$

The term inside the brackets is the maximum of n χ^2-squared distributed random variables. Thus, for all $n \in \mathbb{N}^\star$,

$$\mathbb{E}\left[\max_{1\leq i\leq n}\left(\frac{\varepsilon_i}{\sigma}\right)^2\right] \leq 1 + 4\log n,$$

[see, e.g., Chapter 1 in Boucheron et al., 2013]. Therefore,

$$R(m_{M,n}) - R(m_{\infty,n}) \leq \frac{8}{M} \times \left(\|m\|_\infty^2 + \sigma^2(1+4\log n)\right).$$

3.5.4 Proof of Theorem 3.4 and Corollary 3.1

The proof of Theorem 3.4 is based on Stone's theorem which is recalled here.

Stone's theorem [1977]. *Assume that the following conditions are satisfied for every distribution of* $\mathbf{X}$*:*

(i) There is a constant c such that for every non negative measurable function f satisfying $\mathbb{E}f(\mathbf{X}) < \infty$ and any n,

$$\mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(\mathbf{X})f(\mathbf{X}_i)\right) \leq c\,\mathbb{E}\left(f(\mathbf{X})\right).$$

(ii) There is a $D > 1$ such that, for all n,

$$\mathbb{P}\left(\sum_{i=1}^{n} W_{ni}(\mathbf{X}) < D\right) = 1.$$

(iii) For all $a > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{n} W_{ni}(\mathbf{X})\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|>a}\right) = 0.$$

(iv) The sum of weights satisfies

$$\sum_{i=1}^{n} W_{ni}(\mathbf{X}) \underset{n\to\infty}{\to} 1 \quad \textit{in probability.}$$

(v)

$$\lim_{n\to\infty} \mathbb{E}\left(\max_{1\le i\le n} W_{ni}(\mathbf{X})\right) = 0.$$

Then the corresponding regression function estimate m_n is universally $\mathbb{L}^2$ consistent, that is,

$$\lim_{n\to\infty} \mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0,$$

for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$.

Proof of Theorem 3.4. We check the assumptions of Stone's theorem. For every non negative measurable function f satisfying $\mathbb{E}f(\mathbf{X}) < \infty$ and for any n, almost surely,

$$\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\left(\sum_{i=1}^{n} W_{ni}(\mathbf{X},\Theta) f(\mathbf{X}_i)\right) \le \mathbb{E}_{\mathbf{X}}\left(f(\mathbf{X})\right),$$

where

$$W_{ni}(\mathbf{X},\Theta) = \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}}{N_n(\mathbf{X},\Theta)}$$

are the weights of the random tree $\mathcal{T}_n(\Theta)$ [see the proof of Theorem 4.2 in Györfi et al., 2002]. Taking expectation with respect to Θ from both sides, we have

$$\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\left(\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) f(\mathbf{X}_i)\right) \le \mathbb{E}_{\mathbf{X}}\left(f(\mathbf{X})\right),$$

which proves the first condition of Stone's theorem.

According to the definition of random forest weights W_{ni}^{∞}, since $\sum_{i=1}^{n} W_{ni}(\mathbf{X},\Theta) \le 1$ almost surely, we have

$$\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) = \mathbb{E}_{\Theta}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X},\Theta)\right] \le 1.$$

To check condition (iii), note that, for all $a > 0$,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X})\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty>a}\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty>a}\right] \\ &= \mathbb{E}\Big[\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty>a} \\ &\qquad \times \mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))\geq a/2}\Big],\end{aligned}$$

because $\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty>a}\mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))<a/2} = 0$. Thus,

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X})\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty>a}\Big] &\leq \mathbb{E}\Big[\mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))\geq a/2} \\ &\qquad \times \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty>a}\Big] \\ &\leq \mathbb{P}\Big[\mathrm{diam}(A_n(\mathbf{X},\Theta)) \geq a/2\Big],\end{aligned}$$

which tends to zero, as $n \to \infty$, by assumption.

To prove assumption (iv), we follow the arguments developed by Biau et al. [2008]. For completeness, these arguments are recalled here. Let us consider the partition associated with the random tree $\mathcal{T}_n(\Theta)$. By definition, this partition has 2^k cells, denoted by $A_1, \ldots, A_{2^k}$. For $1 \leq i \leq 2^k$, let N_i be the number of points among $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ falling into A_i. Finally, set $\mathcal{S} = \{\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Since these points are independent and identically distributed, fixing the set $\mathcal{S}$ (but not the order of the points) and Θ, the probability that $\mathbf{X}$ falls in the i-th cell is $N_i/(n+1)$. Thus, for every fixed $t > 0$,

$$\begin{aligned}\mathbb{P}\Big[N_n(\mathbf{X},\Theta) < t\Big] &= \mathbb{E}\Big[\mathbb{P}\Big[N_n(\mathbf{X},\Theta) < t\Big|\mathcal{S},\Theta\Big]\Big] \\ &= \mathbb{E}\left[\sum_{i:N_i<t+1} \frac{N_i}{n+1}\right] \\ &\leq \frac{2^k}{n+1}t.\end{aligned}$$

Thus, by assumption, $N_n(\mathbf{X},\Theta) \to \infty$ in probability, as $n \to \infty$. Consequently, observe that

$$\begin{aligned}\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) &= \mathbb{E}_\Theta\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X},\Theta)\right] \\ &= \mathbb{E}_\Theta\Big[\mathbb{1}_{N_n(\mathbf{X},\Theta)\neq 0}\Big] \\ &= \mathbb{P}_\Theta\left[N_n(\mathbf{X},\Theta) \neq 0\right] \\ &\to 1 \quad \text{as } n \to \infty.\end{aligned}$$

At last, to prove (v), note that,

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}^{\infty}(\mathbf{X})\right] &\leq \mathbb{E}\left[\max_{1\leq i\leq n} \frac{\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)}}{N_n(\mathbf{X},\Theta)}\right]\\ &\leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(\mathbf{X},\Theta)>0}}{N_n(\mathbf{X},\Theta)}\right]\\ &\to 0 \quad \text{as } n\to\infty,\end{aligned}$$

since $N_n(\mathbf{X},\Theta)\to\infty$ in probability, as $n\to\infty$. □

Proof of Corollary 3.1. We check conditions of Theorem 3.4. Let us denote by $V_{nj}(\mathbf{X},\Theta)$ the length of the j-th side of the cell containing $\mathbf{X}$ and $K_{nj}(\mathbf{X},\Theta)$ the number of times the cell containing $\mathbf{X}$ is cut along the j-coordinate. Note that, if $U_1,\dots,U_n$ are independent uniform on $[0,1]$,

$$\begin{aligned}\mathbb{E}\left[V_{nj}(\mathbf{X},\Theta)\right] &\leq \mathbb{E}\left[\mathbb{E}\left[\prod_{l=1}^{K_{nj}(\mathbf{X},\Theta)} \max(U_i,1-U_i)|K_{nj}(\mathbf{X},\Theta)\right]\right]\\ &= \mathbb{E}\Big[\Big[\mathbb{E}\big[\max(U_1,1-U_1)\big]\Big]^{K_{nj}(\mathbf{X},\Theta)}\Big]\\ &= \mathbb{E}\left[\left(\frac{3}{4}\right)^{K_{nj}(\mathbf{X},\Theta)}\right].\end{aligned}$$

Since $K_{nj}(\mathbf{X},\Theta)$ is distributed as a binomial $\mathcal{B}(k_n,1/p)$, $K_{nj}(\mathbf{X},\Theta)\to+\infty$ in probability, as n tends to infinity. Thus $\mathbb{E}\left[V_{nj}(\mathbf{X},\Theta)\right]\to 0$ as $n\to\infty$. □

3.5.5 Proof of Theorem 3.5

Consider a theoretical q quantile tree where cuts are made similarly as in the q quantile tree (defined in Algorithm 3) but by selecting $q'\in[1-q,q]$ and by performing the cut at the q' theoretical quantile (instead of empirical one). The tree is then stopped at level k, where $k\in\mathbb{N}$ is a parameter to be chosen later. Denote by $A_k^\star(\mathbf{X},\Theta)$ the cell of the theoretical q quantile tree of level k containing $\mathbf{X}$ and built with the randomness Θ. Finally, we let $\mathbf{d}_k^\star = (d_1^\star(\mathbf{X},\Theta),\dots,d_k^\star(\mathbf{X},\Theta))$ be the k cuts used to construct the cell $A_k^\star(\mathbf{X},\Theta)$.

To prove Theorem 3.5, we need the following lemma which states that the cell diameter of a theoretical q quantile tree tends to zero.

Lemma 3.2. *Assume that $\mathbf{X}$ has a density over $[0,1]^p$, with respect to the Lebesgue measure. Thus, for all $q\in[1/2,1)$, the theoretical q quantile tree defined above satisfies, for all γ,*

$$\mathbb{P}\big[diam(A_k^\star(\mathbf{X},\Theta))>\gamma\big] \underset{k\to\infty}{\to} 0.$$

Proof of Lemma 3.2. Set $q\in[1/2,1)$ and consider a theoretical q quantile tree. For all $A\subset[0,1]^p$, let

$$\mu(A)=\int_A f d\nu,$$

where ν is the Lebesgue measure, and f the density of $\mathbf{X}$. Take $z \in [0,1]$, $\ell \in \{1,\dots,p\}$ and let Δ be the hyperplane such that $\Delta = \{\mathbf{x} : x^{(\ell)} = z\}$. At last, we denote by $\mathrm{D} = \{A : A \cap \Delta \neq \emptyset\}$ the set of cells of the theoretical q quantile tree that have a non-empty intersection with Δ.

If a cell $A_k^\star(\mathbf{X},\Theta)$ belongs to D, then:

Case 1 Either the next split in $A_k^\star(\mathbf{X},\Theta)$ is performed along the ℓ-th coordinate and, in that case, one of the two resulting cell has an empty intersection with Δ. Note that the measure of this cell is, at least, $(1-q)\mu(A_k^\star(\mathbf{X},\Theta))$.

Case 2 Or the next split is performed along the j-th coordinate (with $j \neq \ell$) and, in that case, the two resulting cells have a non-empty intersection with Δ.

Since the splitting directions are chosen uniformly over $\{1,\dots,p\}$, for each cell **Case 1** occurs with probability $1/p$ and **Case 2** with probability $1-1/p$. Let $j_k(\mathbf{X},\Theta)$ be the random variable equals to the coordinate along which the split in the cell $A_{k-1}^\star(\mathbf{X},\Theta)$ is performed. Thus,

$$\begin{aligned}
&\mathbb{P}\left[A_{k+1}^\star(\mathbf{X},\Theta) \in \mathrm{D}\right] \\
&\leq \mathbb{E}\Big[\mathbb{P}\left[A_{k+1}^\star(\mathbf{X},\Theta) \in \mathrm{D}\right] \Big| j_{k+1}(\mathbf{X},\Theta)\Big] \\
&\leq \mathbb{E}\Big[\mathbb{P}\left[A_k^\star(\mathbf{X},\Theta) \in \mathrm{D}\right](1-q)\mathbb{1}_{j_{k+1}(\mathbf{X},\Theta)=\ell} \\
&\qquad + \mathbb{P}\left[A_k^\star(\mathbf{X},\Theta) \in \mathrm{D}\right]\mathbb{1}_{j_{k+1}(\mathbf{X},\Theta)\neq\ell}\Big] \\
&\leq \mathbb{P}\left[A_k^\star(\mathbf{X},\Theta) \in \mathrm{D}\right] \\
&\qquad \times \Big((1-q)\mathbb{P}\left[j_{k+1}(\mathbf{X},\Theta) = \ell\right] + \mathbb{P}\left[j_{k+1}(\mathbf{X},\Theta) \neq \ell\right]\Big) \\
&\leq \left(1 - \frac{q}{p}\right)\mathbb{P}\left[A_k^\star(\mathbf{X},\Theta) \in \mathrm{D}\right].
\end{aligned}$$

Consequently, for all k,

$$\mathbb{P}\left[A_{k+1}^\star(\mathbf{X},\Theta) \in \mathrm{D}\right] \leq \left(1 - \frac{q}{p}\right)^k \mathbb{P}\left[A_k^\star(\mathbf{X},\Theta) \in \mathrm{D}\right], \tag{3.16}$$

that is

$$\mathbb{P}\left[A_k^\star(\mathbf{X},\Theta) \in \mathrm{D}\right] \underset{k\to\infty}{\to} 0. \tag{3.17}$$

To finish the proof, take $\varepsilon > 0$ and consider a $\varepsilon \times \ldots \times \varepsilon$ grid. Within a grid cell, all points are distant from, at most, $\varepsilon p^{1/2}$. Thus, if a cell A of the median tree is contained in a grid cell, it satisfies

$$\mathrm{diam}(A) \leq \varepsilon p^{1/2}.$$

Consider the collection of hyperplane that correspond to the grid, that is all hyperplanes of the form $\{x : x^{(\ell)} = j\varepsilon\}$ for $\ell \in \{1,\dots,p\}$ and $j \in \{0,\dots,\lfloor 1/\varepsilon \rfloor\}$. Denote by Δ_{grid} the collection of these hyperplanes. Since the number of hyperplanes is finite, according to (3.17), we have

$$\mathbb{P}\left[\mathrm{diam}(A_k^\star(\mathbf{X},\Theta)) \geq \varepsilon p^{1/2}\right] \leq \mathbb{P}\left[A_k^\star(\mathbf{X},\Theta) \cap \Delta_{grid} \neq \emptyset\right] \underset{k\to\infty}{\to} 0,$$

which concludes the proof. □

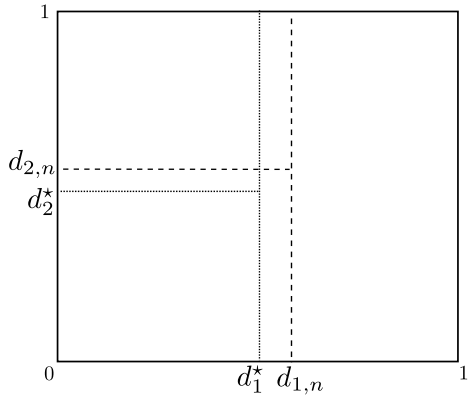

Figure 3.3: Respective positions of theoretical and empirical splits in a median tree.

Recall that $A_n(\mathbf{X}, \Theta)$ is the cell of the q quantile tree containing $\mathbf{X}$. Similarly, $A_{k,n}(\mathbf{X}, \Theta)$ is the cell of the q quantile tree containing $\mathbf{X}$ where only the first k cuts ($k \in \mathbb{N}^\star$) are performed . We denote by $\mathbf{d}_{k,n} = (d_{1,n}(\mathbf{X}, \Theta), \ldots, d_{k,n}(\mathbf{X}, \Theta))$ the k cuts used to construct the cell $A_{k,n}(\mathbf{X}, \Theta)$.

Lemma 3.3. *Assume that* $\mathbf{X}$ *has a density over* $[0,1]^p$*, with respect to the Lebesgue measure. Thus, for all* $k \in \mathbb{N}$*, a.s.*

$$\|\mathbf{d}_{k,n} - \mathbf{d}_k^\star\|_\infty \underset{n\to\infty}{\to} 0.$$

Proof of Lemma 3.3. To keep the argument simple, we fix $\mathbf{X} \in [0,1]^p$ and assume that the first and second splits are performed at the empirical median along the first (resp. second) coordinate. Since $\mathbf{X}$ and Θ are fixed, we omit the dependency in $\mathbf{X}$ and Θ in the rest of the proof. Let $d_{1,n}$ (resp. $d_{2,n}$) be the position of the first (resp. second) splits along the first (resp. second) axis. We denote by $d_1^\star$ (resp. $d_2^\star$) the position of the theoretical median of the distribution (see Figure 3.3).

Fix $\varepsilon > 0$. Since the X_i's are i.i.d., the empirical median tends to the theoretical median almost surely. With our notation, a.s., $d_{1,n} \to d_1^\star$, as n tends to infinity. Therefore, Lemma 3.3 holds for $k = 1$. We now prove Lemma 3.3 for $k = 2$. To this end, we define, for all $0 \le a < b \le 1$, the subset $H_{a,b}$ of the cell $A_1^\star$ by

$$H_{a,b} = [0,1] \times [a,b] \times [0,1] \times \ldots \times [0,1] \cap A_1^\star.$$

Let $\alpha = \min(\mu(H_{d_2^\star-\varepsilon,d_2^\star}), \mu(H_{d_2^\star,d_2^\star+\varepsilon}))$. Denote by $d_{2,n}(d_1^\star)$ the empirical median of data points falling into the cell $A_1^\star$. Since $\mathbf{X}$ has a density on $[0,1]^p$, one can find ε_1 such that, for all n large enough, a.s.,

$$\begin{cases} |d_{2,n}(d_1^\star) - d_2^\star| \le \varepsilon_1 \\ \min(\mu(H_{d_2^\star-\varepsilon_1,d_2^\star}), \mu(H_{d_2^\star,d_2^\star+\varepsilon_1})) \le \alpha/100. \end{cases}$$

By the same argument, one can find ε_2 such that, for all n large enough, a.s.,

$$\begin{cases} |d_{1,n} - d_1^\star| \leq \varepsilon_2 \\ \min(\mu(H_{d_1^\star-\varepsilon_2,d_1^\star}), \mu(H_{d_1^\star,d_1^\star+\varepsilon_2})) \leq \alpha/100. \end{cases}$$

A direct consequence of the law of the iterated logarithm applied to cumulative distribution function is that, for all n large enough, a.s.,

$$\max\left(N_n(H_{d_2^\star-\varepsilon_1,d_2^\star}), N_n(H_{d_2^\star,d_2^\star+\varepsilon_1})\right) \leq 0.02\alpha n, \tag{3.18}$$

$$\max\left(N_n(H_{d_1^\star-\varepsilon_2,d_1^\star}), N_n(H_{d_1^\star,d_1^\star+\varepsilon_2})\right) \leq 0.02\alpha n, \tag{3.19}$$

$$\text{and} \quad \min\left(N_n(H_{d_2^\star-\varepsilon,d_2^\star-\varepsilon_1}), N_n(H_{d_2^\star+\varepsilon_1,d_2^\star+\varepsilon})\right) \geq 0.98\alpha n. \tag{3.20}$$

The empirical median in the cell $A_{1,n}$ is given by $X^{(2)}_{(\lfloor N_n(A_{1,n})/2 \rfloor)}$, where the $\mathbf{X}_i$'s are sorted along the second coordinate. According to (3.19), the cell $A_1^{(\star)}$ contains at most $N_n(A_{1,n}) + 0.02\alpha n$. Therefore, the empirical median $d_{2,n}(d_1^\star)$ in the cell $A_1^\star$ is at most $X^{(2)}_{(\lfloor (N_n(A_{1,n})+0.02\alpha n)/2 \rfloor)}$. Thus, according to (3.18) and (3.20),

$$d_{2,n} \leq X^{(2)}_{(\lfloor (N_n(A_{1,n})+0.02\alpha n)/2 \rfloor)} \leq d_2^\star + \varepsilon.$$

Similarly, one has

$$d_{2,n} \geq X^{(2)}_{(\lfloor (N_n(A_{1,n})-0.02\alpha n)/2 \rfloor)} \leq d_2^\star - \varepsilon.$$

Consequently, for all n large enough, a.s., $|d_{2,n} - d_2^\star| \leq \varepsilon$. The extension for arbitrary k is straightforward. □

Lemma 3.4. *Assume that* $\mathbf{X}$ *has a density over* $[0,1]^p$*, with respect to the Lebesgue measure. Thus, for all* $q \in [1/2, 1)$*, the theoretical* q *quantile tree defined above satisfies, for all* γ*,*

$$\mathbb{P}\left[diam(A_n(\mathbf{X}, \Theta)) > \gamma\right] \underset{n\to\infty}{\to} 0.$$

Proof. Now, consider the empirical q quantile tree as defined in Algorithm 3 but stopped at level k. Thus, for n large enough, at each step of the algorithm, q' is selected in $[1-q, q]$. Set $\varepsilon, \gamma > 0$. By Lemma 3.2, there exists $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$,

$$\mathbb{P}\left[\text{diam}(A_k(\mathbf{X}, \Theta)) > \gamma\right] \leq \varepsilon.$$

Thus, according to Lemma 3.3, for all n large enough, a.s.,

$$\mathbb{P}\left[\text{diam}(A_{k_0,n}(\mathbf{X}, \Theta)) > \gamma/2\right] \leq \varepsilon.$$

Since, for all n large enough, a.s.,

$$\text{diam}(A_{k_0,n}(\mathbf{X}, \Theta)) \geq \text{diam}(A_n(\mathbf{X}, \Theta)),$$

the proof is complete.

□

Proof of Theorem 3.5. We check the conditions of Stone's theorem. Condition (i) is satisfied since the regression function is uniformly continuous and $\mathrm{Var}[Y|\mathbf{X}=\mathbf{x}] \leq \sigma^2$ [see remark after Stone theorem in Györfi et al., 2002].

Condition (ii) is always satisfied for random trees. Condition (iii) is verified since

$$\mathbb{P}\left[\mathrm{diam}(A_n(\mathbf{X},\Theta)) > \gamma\right] \underset{n\to\infty}{\to} 0,$$

according to Lemma 3.4.

Since each cell contains exactly one data point,

$$\begin{aligned}\sum_{i=1}^{n} W_{ni}(x) =& \sum_{i=1}^{n} \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}}{N_n(\mathbf{X},\Theta)}\right] \\ =& \mathbb{E}_\Theta\left[\frac{1}{N_n(\mathbf{X},\Theta)}\sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}\right] \\ =& 1.\end{aligned}$$

Thus, conditions (iv) of Stone theorem is satisfied.

To check (v), observe that in the subsampling step, there are exactly $\binom{a_n-1}{n-1}$ choices to pick a fixed observation $\mathbf{X}_i$. Since $\mathbf{x}$ and $\mathbf{X}_i$ belong to the same cell only if $\mathbf{X}_i$ is selected in the subsampling step, we see that

$$\mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right] \leq \frac{\binom{a_n-1}{n-1}}{\binom{a_n}{n}} = \frac{a_n}{n}.$$

So,

$$\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(\mathbf{X})\right] \leq \mathbb{E}\left[\max_{1\leq i\leq n} \mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right]\right] \leq \frac{a_n}{n},$$

which tends to zero by assumption. □

3.5.6 Proofs of Technical Lemmas 3.1 and 3.2

Technical Lemma 3.1. *Take $k \in \mathbb{N}$ and consider a uniform random forest where each tree is stopped at level k. For all $\mathbf{x}, \mathbf{z} \in [0,1]^p$, its connection function satisfies*

$$K_k(0, |\mathbf{x}-\mathbf{z}|) \leq K_k(\mathbf{x}, \mathbf{z}),$$

where $|\mathbf{x}-\mathbf{z}| = (|x_1 - z_1|, \ldots, |x_d - z_d|)$.

Proof. Take $x, z \in [0,1]$. Without loss of generality, one can assume that $x < z$ and let $\mu = z - x$. Consider the following two configurations.

For any $k \in \mathbb{N}^\star$, we let $\mathbf{d}_k = (d_1, \ldots, d_k)$ (resp. $\mathbf{d}'_k = (d'_1, \ldots, d'_k)$) be k consecutive cuts in configuration 1 (resp. in configuration 2). We denote by $\mathcal{A}_k$ (resp. $\mathcal{A}'_k$) the set where $\mathbf{d}_k$ (resp. $\mathbf{d}'_k$) belong.

We show that for all $k \in \mathbb{N}^\star$, there exists a coupling between $\mathcal{A}_k$ and $\mathcal{A}'_k$ satisfying the following property: any k-tuple $\mathbf{d}_k$ is associated with a k-tuple $\mathbf{d}'_k$ such that

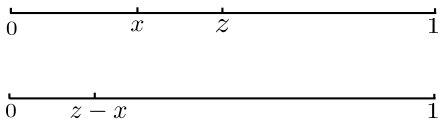

Figure 3.4: Scheme of configuration 1 (at the top) and 2 (at the bottom).

1. if $\mathbf{d}_k$ separates $[x, z]$ then $\mathbf{d}'_k$ separates $[0, z - x]$,
2. if $\mathbf{d}_k$ does not separate $[x, z]$ and $\mathbf{d}'_k$ does not separate $[0, z-x]$, then the length of the cell containing $[x, z]$ built with $\mathbf{d}_k$ is higher than the one containing $[0, z - x]$ built with $\mathbf{d}'_k$.

We call $\mathcal{H}_k$ this property. We now proceed by induction. For $k = 1$, we use the function g to map $\mathcal{A}_1$ into $\mathcal{A}'_1$ such that:

$$g_1(u) = \begin{cases} u & \text{if } u > z \\ z - u & \text{if } u \leq z \end{cases}$$

Thus, for any $d_1 \in \mathcal{A}_1$, if d_1 separates $[x, z]$, then $d'_1 = g_1(d_1)$ separates $[0, z - x]$. Besides, the length of the cell containing $[x, z]$ designed with the cut d_1 is higher than that of the cell containing $[0, z - x]$ designed with the cut d'_1. Consequently, $\mathcal{H}_1$ is true.

Now, take $k > 1$ and assume that $\mathcal{H}_k$ is true. Consequently, if $\mathbf{d}_k$ separates $[x, z]$ then $g_k(\mathbf{d}_k)$ separates $[0, z-x]$. In that case, $\mathbf{d}_{k+1}$ separates $[x, z]$ and $g_{k+1}(\mathbf{d}_{k+1})$ separates $[0, z-x]$. Thus, in the rest of the proof, we assume that $\mathbf{d}_k$ does not separate $[x, z]$ and $g_k(\mathbf{d}_k)$ does not separate $[0, z - x]$. Let $[a_k, b_k]$ be the cell containing $[x, z]$ built with cuts $\mathbf{d}_k$. Since the problem is invariant by translation, we assume, without loss of generality, that $[a_k, b_k] = [0, \delta_k]$, where $\delta_k = b_k - a_k$ and $[x, z] = [x_k, x_k + \mu]$ (see Figure 3.5).

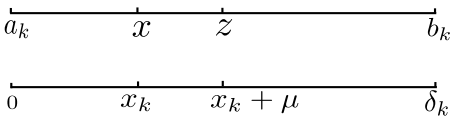

Figure 3.5: Configuration 1a (at the top) and 1b (at the bottom).

In addition, according to $\mathcal{H}_k$, the length of the cell built with $\mathbf{d}_k$ is higher than the one built with $\mathbf{d}'_k$. Thus, one can find $\lambda \in (0, 1)$ such that $d'_k = \lambda\delta_k$. This is summarized in Figure 3.6.

Thus, one can map $[0, \delta_k]$ into $[0, \lambda\delta_k]$ with g_{k+1} defined as

$$g_{k+1}(u) = \begin{cases} \lambda u & \text{if } u > x_k + \mu \\ \lambda(x_k + \mu - u) & \text{if } u \leq x_k + \mu \end{cases}$$

Note that, for all d_{k+1}, the length of the cell containing $[x_k, x_k + \mu]$ designed with the cut d_{k+1} (configuration 1b) is bigger than the length of the cell containing $[0, \mu]$ designed with the cut

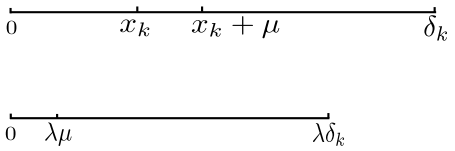

Figure 3.6: Configuration 1b (at the top) and 2b (at the bottom).

$d'_{k+1} = g_{k+1}(d_{k+1})$ (configuration 2b). Besides, if $d_{k+1} \in [x_k, x_k + \mu]$ then $g_{k+1}(d_{k+1}) \in [0, \mu]$. Consequently, the set of functions $g_1, \ldots, g_{k+1}$ induce a mapping of $\mathcal{A}_{k+1}$ into $\mathcal{A}'_{k+1}$ such that $\mathcal{H}_{k+1}$ holds. Thus, Technical Lemma 3.1 holds for $p = 1$.

To address the case where $p > 1$, note that

$$\begin{aligned} K_k(\mathbf{x}, \mathbf{z}) &= \sum_{\substack{k_1,\ldots,k_p \\ \sum_{j=1}^p k_j = k}} \frac{k!}{k_1! \ldots k_p!} \left(\frac{1}{p}\right)^k \prod_{m=1}^p K_{k_m}(x_m, z_m) \\ &\geq \sum_{\substack{k_1,\ldots,k_p \\ \sum_{j=1}^p k_j = k}} \frac{k!}{k_1! \ldots k_p!} \left(\frac{1}{p}\right)^k \prod_{m=1}^p K_{k_m}(0, |z_m - x_m|) \\ &\geq K_k(0, |\mathbf{z} - \mathbf{x}|), \end{aligned}$$

which concludes the proof.

□

Technical Lemma 3.2. *Take $k \in \mathbb{N}$ and consider a uniform random forest where each tree is stopped at level k. For all $x \in [0, 1]$, its connection function $K_k(0, x)$ satisfies*

$$K_k(0, x) = 1 - x \sum_{j=0}^{k-1} \frac{(-\ln x)^j}{j!},$$

with the notational convention that the last sum is zero if $k = 0$.

Proof of Technical Lemma 3.2. The result is clear for $k = 0$. Thus, set $k \in \mathbb{N}^\star$ and consider a uniform random forest where each tree is stopped at level k. Since the result is clear for $x = 0$, take $x \in]0, 1]$ and let $I = [0, x]$. Thus

$$\begin{aligned} K_k(0, x) &= \mathbb{P}\left[0 \overset{\Theta}{\underset{k \text{ cuts}}{\leftrightarrow}} x\right] \\ &= \int_{z_1 \notin I} \int_{z_2 \notin I} \ldots \int_{z_k \notin I} \nu(\mathrm{d}z_k | z_{k-1}) \nu(\mathrm{d}z_{k-1} | z_{k-2}) \ldots \nu(\mathrm{d}z_2 | z_1) \nu(\mathrm{d}z_1), \end{aligned}$$

where $z_1, \ldots, z_k$ are the positions of the k cuts (see Figure 3.7).

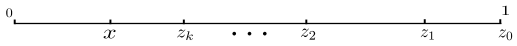

Figure 3.7: Positions of cuts $z_1, \ldots, z_k$ and x with $d = 1$

We prove by induction that, for every integer ℓ,

$$\int_{z_{k-\ell} \notin I} \ldots \int_{z_k \notin I} \nu(\mathrm{d}z_k | z_{k-1}) \ldots \nu(\mathrm{d}z_{k-\ell} | z_{k-\ell-1}) = 1 - \frac{x}{z_{k-\ell-1}} \left(\sum_{j=0}^{\ell} \frac{[\ln(z_{k-\ell-1}/x)]^j}{j!} \right).$$

Denote by $\mathcal{H}_\ell$ this property. Since, given z_{k-1}, z_k is uniformly distributed over $[0, z_{k-1}]$, we have

$$\int_{z_k \notin I} \nu(\mathrm{d}z_k | z_{k-1}) = 1 - \frac{x}{z_{k-1}}.$$

Thus $\mathcal{H}_0$ is true. Now, fix $\ell > 0$ and assume that $\mathcal{H}_\ell$ is true. Let $u = z_{k-\ell-1}/x$. Thus, integrating both sides of $\mathcal{H}_\ell$, we deduce,

$$\begin{aligned}
&\int_{z_{k-\ell-1} \notin I} \int_{z_{k-\ell} \notin I} \ldots \int_{z_k \notin I} \nu(\mathrm{d}z_k | z_{k-1}) \ldots \nu(\mathrm{d}z_{k-\ell} | z_{k-\ell-1}) \nu(\mathrm{d}z_{k-\ell-1} | z_{k-\ell-2}) \\
&= \int_{z_{k-\ell-1} \notin I} \left[1 - \frac{x}{z_{k-\ell-1}} \left(\sum_{j=0}^{\ell} \frac{[\ln(z_{k-\ell-1}/x)]^j}{j!} \right) \right] \nu(\mathrm{d}z_{k-\ell-1} | z_{k-\ell-2}) \\
&= \int_x^{z_{k-\ell-2}} \left[1 - \frac{x}{z_{k-\ell-1}} \left(\sum_{j=0}^{\ell} \frac{[\ln(z_{k-\ell-1}/x)]^j}{j!} \right) \right] \frac{\mathrm{d}z_{k-\ell-1}}{z_{k-\ell-2}} \\
&= \frac{x}{z_{k-\ell-2}} \int_1^{z_{k-\ell-2}/x} \left[1 - \frac{1}{u} \left(\sum_{j=0}^{\ell} \frac{[\ln(u)]^j}{j!} \right) \right] \mathrm{d}u.
\end{aligned}$$

Using integration by parts on the last term, we conclude that $\mathcal{H}_{\ell+1}$ is true. Thus, for all $\ell > 0$, $\mathcal{H}_\ell$ is verified. Finally, using $\mathcal{H}_{k-1}$ and the fact that $z_0 = 1$, we conclude the proof. □

Chapter 4

Random forests and kernel methods

Abstract Random forests are ensemble methods which grow trees as base learners and combine their predictions by averaging. Random forests are known for their good practical performance, particularly in high dimensional settings. On the theoretical side, several studies highlight the potentially fruitful connection between random forests and kernel methods. In this paper, we work out in full details this connection. In particular, we show that by slightly modifying their definition, random forests can be rewritten as kernel methods (called KeRF for Kernel based on Random Forests) which are more interpretable and easier to analyze. Explicit expressions of KeRF estimates for some specific random forest models are given, together with upper bounds on their rate of consistency. We also show empirically that KeRF estimates compare favourably to random forest estimates.

We would like to thank Arthur Pajot for his great help in the implementation of KeRF estimates.

Contents

4.1 Introduction

Random forests are a class of learning algorithms used to solve pattern recognition problems. As ensemble methods, they grow many trees as base learners and aggregate them to predict. Growing many different trees from a single data set requires to randomize the tree building process by, for example, sampling the data set. Thus, there exists a variety of random forests, depending on how trees are built and how the randomness is introduced in the tree building process.

One of the most popular random forests is that of Breiman [2001] which grows trees based on CART procedure [Classification and Regression Trees, Breiman et al., 1984] and randomizes both the training set and the splitting directions. Breiman's [2001] random forests have been under active investigation during the last decade mainly because of their good practical performance and their ability to handle high dimensional data sets. Moreover, they are easy to run since they only depend on few parameters which are easily tunable [Liaw and Wiener, 2002, Genuer et al., 2008]. They are acknowledged to be state-of-the-art methods in fields such as genomics [Qi, 2012] and pattern recognition [Rogez et al., 2008], just to name a few.

However, even if random forests are known to perform well in many contexts, little is known about their mathematical properties. Indeed, most authors study forests whose construction does not depend on the data set. Although, consistency of such simplified models has been addressed in the literature [e.g., Biau et al., 2008, Ishwaran and Kogalur, 2010, Denil et al., 2013], these results do not adapt to Breiman's forests whose construction strongly depends on the whole training set. The latest attempts to study the original algorithm are by Mentch and Hooker [2014a] and Wager [2014] who prove its asymptotic normality or by Scornet et al. [2015b] who prove its consistency under appropriate assumptions.

Despite these works, several properties of random forests still remain unexplained. A promising way for understanding their complex mechanisms is to study the connection between forests and kernel estimates, that is estimates m_n which take the form

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_k(\mathbf{X}_i, \mathbf{x})}{\sum_{i=1}^n K_k(\mathbf{X}_i, \mathbf{x})}, \tag{4.1}$$

where $\{(\mathbf{X}_i, Y_i) : 1 \leq i \leq n\}$ is the training set, $(K_k)_k$ is a sequence of kernel functions, and k ($k \in \mathbb{N}$) is a parameter to be tuned. Unlike the most used Nadaraya-Watson kernels [Nadaraya, 1964, Watson, 1964] which satisfy a homogeneous property of the form $K_h(\mathbf{X}_i, \mathbf{x}) = K((\mathbf{x} - \mathbf{X}_i)/h)$, kernels K_k are not necessarily of this form. Therefore, the analysis of kernel estimates defined by (4.1) turns out to be more complicated and cannot be based on general results regarding Nadaraya-Watson kernels.

Breiman [2000a] was the first to notice the link between forest and kernel methods, a link which was later formalized by Geurts et al. [2006]. On the practical side, Davies and Ghahramani [2014] highlight the fact that a specific kernel based on random forests can empirically outperform state-of-the-art kernel methods. Another approach is taken by Lin and Jeon [2006] who establish the connection between random forests and adaptive nearest neighbor, implying that random forests can be seen as adaptive kernel estimates [see also Biau and Devroye, 2010]. The latest study is by Arlot and Genuer [2014] who show that a specific random forest can be written as a kernel estimate and who exhibit rates of consistency. However, despite these works, the literature is relatively sparse regarding the link between forests and kernel methods.

Our objective in the present paper is to prove that a slight modification of random forest procedures have explicit and simple interpretations in terms of kernel methods. Thus, the resulting kernel based on random forest (called KeRF in the rest of the paper) estimates are more amenable to mathematical analysis. They also appear to be empirically as accurate as random forest estimates. To theoretically support these results, we also make explicit the expression of some KeRF. We prove upper bounds on their rates of consistency, which compare favorably to the existing ones.

The paper is organized as follows. Section 2 is devoted to notations and to the definition of KeRF estimates. The link between KeRF estimates and random forest estimates is made explicit in Section 3. In Section 4, two KeRF estimates are presented and their consistency is proved along with their rate of consistency. Section 5 contains experiments that highlight the good performance of KeRF compared to their random forests counterparts. Proofs are postponed to Section 6.

4.2 Notations and first definitions

4.2.1 Notations

Throughout the paper, we assume to be given a training sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ of $[0,1]^d \times \mathbb{R}$-valued independent random variables distributed as the independent prototype pair $(\mathbf{X}, Y)$, where $\mathbb{E}[Y^2] < \infty$. We aim at predicting the response Y, associated with the random variable $\mathbf{X}$, by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. In this context, we use infinite random forests (see the definition below) to build an estimate $m_{\infty,n} : [0,1]^d \to \mathbb{R}$ of m, based on the data set $\mathcal{D}_n$.

A random forest is a collection of M randomized regression trees [for an overview on tree construction, see e.g., Chapter 20 in Györfi et al., 2002]. For the j-th tree in the family, the predicted value at point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}, \Theta_j)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ, independent of the sample $\mathcal{D}_n$. This random variable can be used to sample the training set or to select the candidate directions or positions for splitting. The trees are combined to form the finite forest estimate

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}, \Theta_j).$$

By the law of large numbers, for all $\mathbf{x} \in [0,1]^d$, almost surely, the finite forest estimate tends to the infinite forest estimate

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}, \Theta) \right],$$

where $\mathbb{E}_\Theta$ denotes the expectation with respect to Θ, conditionally on $\mathcal{D}_n$.

As mentioned above, there is a large variety of forests, depending on how trees are grown and how the random variable Θ influences the tree construction. For instance, tree construction can be independent of $\mathcal{D}_n$ [Biau, 2012]. On the other hand, it can depend only on the $\mathbf{X}_i$'s [Biau et al., 2008] or on the whole training set [Cutler and Zhao, 2001, Geurts et al., 2006, Zhu et al., 2012]. Throughout the paper, we use three important types of random forests to exemplify our results: Breiman's, centred and uniform forests. In Breiman's original procedure, splits are performed to minimize the variances within the two resulting cells. The algorithm stops when each cell contains less than a small pre-specified number of points [typically between 1 and 5; see Breiman, 2001, for details]. Centred forests are a simpler procedure which, at each node, uniformly select a coordinate among $\{1, \ldots, d\}$ and performs splits at the center of the cell along the pre-chosen coordinate. The algorithm stops when a full binary tree of level k is built (that is, each cell is cut exactly k times), where $k \in \mathbb{N}$ is a parameter of the algorithm [see Breiman,

2004, for details on the procedure]. Uniform forests are quite similar to centred forests except that once a split direction is chosen, the split is drawn uniformly on the side of the cell, along the preselected coordinate [see, e.g., Arlot and Genuer, 2014].

4.2.2 Kernel based on random forests (KeRF)

To be more specific, random forest estimates satisfy, for all $\mathbf{x} \in [0,1]^d$,

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} \frac{Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}}{N_n(\mathbf{x}, \Theta_j)} \right),$$

where $A_n(\mathbf{x}, \Theta_j)$ is the cell containing $\mathbf{x}$, designed with randomness Θ_j and data set $\mathcal{D}_n$, and

$$N_n(\mathbf{x}, \Theta_j) = \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}$$

is the number of data points falling in $A_n(\mathbf{x}, \Theta_j)$. Note that, the weights $W_{i,j,n}(\mathbf{x})$ of each observation Y_i defined by

$$W_{i,j,n}(\mathbf{x}) = \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}}{N_n(\mathbf{x}, \Theta_j)}$$

depend on the number of observations $N_n(\mathbf{x}, \Theta_j)$. Thus the contributions of observations that are in cells with a high density of data points are smaller than that of observations which belong to less populated cells. This is particularly true for non adaptive forests (i.e., forests built independently of data) since the number of observations in each cell cannot be controlled. Giving important weights to observations that are in low-density cells can potentially lead to rough estimates. Indeed, as an extreme example, trees of non adaptive forests can contain empty cells which leads to a substantial misestimation (since the prediction in empty cells is set, by default, to zero).

In order to improve the random forest methods and compensate the misestimation induced by random forest weights, a natural idea is to consider KeRF estimates defined, for all $\mathbf{x} \in [0,1]^d$, by

$$\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{\sum_{j=1}^{M} N_n(\mathbf{x}, \Theta_j)} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}. \tag{4.2}$$

Note that $\tilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)$ is equal to the mean of the Y_i's falling in the cells containing $\mathbf{x}$ in the forest. Thus, each observation is weighted by the number of times it appears in the trees of the forests. Consequently, in this setting, an empty cell does not contribute to the prediction.

The proximity between KeRF estimates $\widetilde{m}_{M,n}$ and random forest estimates will be thoroughly discussed in Section 3. As for now, we focus on (4.2) and start by proving that it is indeed a kernel estimate whose expression is given by Proposition 4.1.

Proposition 4.1. *Almost surely, for all* $\mathbf{x} \in [0,1]^d$*, we have*

$$\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{\sum_{i=1}^n Y_i K_{M,n}(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^n K_{M,n}(\mathbf{x}, \mathbf{X}_\ell)}, \tag{4.3}$$

where

$$K_{M,n}(\mathbf{x}, \mathbf{z}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}_{\mathbf{z} \in A_n(\mathbf{x}, \Theta_j)}. \tag{4.4}$$

We call $K_{M,n}$ *the connection function of the* M *finite forest.*

Proposition 4.1 states that KeRF estimates have a more interpretable form than random forest estimates since their kernels are the connection functions of the forests. This connection function can be seen as a geometrical characteristic of the cells in the random forest. Indeed, fixing $\mathbf{X}_i$, the quantity $K_{M,n}(\mathbf{x}, \mathbf{X}_i)$ is nothing but the empirical probability that $\mathbf{X}_i$ and $\mathbf{x}$ are connected (i.e. in the same cell) in the M finite random forest. Thus, the connection function is a natural way to build kernel functions from random forests, a fact that had already been noticed by Breiman [2001]. Note that these kernel functions have the nice property of being positive semi-definite, as proved by Davies and Ghahramani [2014].

A natural question is to ask what happens to KeRF estimates when the number of trees M goes to infinity. To this aim, we define infinite KeRF estimates $\widetilde{m}_{\infty,n}$ by, for all $\mathbf{x}$,

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \lim_{M \to \infty} \widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M). \tag{4.5}$$

In addition, we say that an infinite random forest is discrete (resp. continuous) if its connection function K_n is piecewise constant (resp. continuous). For example, Breiman forests and centred forests are discrete but uniform forests are continuous. Denote by $\mathbb{P}_\Theta$ the probability with respect to Θ, conditionally on $\mathcal{D}_n$. Proposition 4.2 extends the results of Proposition 4.1 to the case of infinite KeRF estimates.

Proposition 4.2. *Consider an infinite discrete or continuous forest. Then, almost surely, for all* $\mathbf{x}, \mathbf{z} \in [0,1]^d$*,*

$$\lim_{M \to \infty} K_{M,n}(\mathbf{x}, \mathbf{z}) = K_n(\mathbf{x}, \mathbf{z}),$$

where

$$K_n(\mathbf{x}, \mathbf{z}) = \mathbb{P}_\Theta \left[\mathbf{z} \in A_n(\mathbf{x}, \Theta) \right].$$

We call K_n *the connection function of the infinite random forest. Thus, for all* $\mathbf{x} \in [0,1]^d$*, one has*

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_n(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^n K_n(\mathbf{x}, \mathbf{X}_\ell)}.$$

This lemma shows that infinite KeRF estimates are kernel estimates with kernel function equal to K_n. Observing that $K_n(\mathbf{x}, \mathbf{z})$ is the probability that $\mathbf{x}$ and $\mathbf{z}$ are connected in the infinite forest, the function K_n characterizes the shape of the cells in the infinite random forest.

Now that we know the expression of KeRF estimates, we are ready to study how close this approximation is to random forest estimates. This link will be further work out in Section 4 for centred and uniform KeRF and empirically studied in Section 5.

4.3 Relation between KeRF and random forests

In this section, we investigate in which cases KeRF and forest estimates are close to each other. To achieve this goal, we will need the following assumption.

(H4.1) (H1) *Fix* $\mathbf{x} \in [0,1]^d$*, and assume that* $Y \geq 0$ *a.s.. Then, one of the following two conditions holds:*

(H1.1) *There exist sequences* $(a_n), (b_n)$ *such that, a.s.,*

$$a_n \leq N_n(\mathbf{x}, \Theta) \leq b_n.$$

(H1.2) *There exist sequences* $(\varepsilon_n), (a_n), (b_n)$ *such that, a.s.,*

- $1 \leq a_n \leq \mathbb{E}_\Theta\left[N_n(\mathbf{x}, \Theta)\right] \leq b_n$,
- $\mathbb{P}_\Theta\Big[a_n \leq N_n(\mathbf{x}, \Theta) \leq b_n\Big] \geq 1 - \varepsilon_n$.

(H1) assumes that the number of points in every cell of the forest can be bounded from above and below. **(H1.1)** holds for finite forests for which the number of points in each cell is controlled almost surely. Typically, **(H1.1)** is verified for adaptive random forests, if the stopping rule is properly chosen. On the other hand, **(H1.2)** holds for infinite forests. Note that the first condition $\mathbb{E}_\Theta\left[N_n(\mathbf{x}, \Theta)\right] \geq 1$ in **(H1.2)** is technical and is true if the level of each tree is tuned appropriately. Several random forests which satisfy **(H1)** are discussed below.

Proposition 4.3 states that finite forest estimate $m_{M,n}$ and finite KeRF estimate $\widetilde{m}_{M,n}$ are close to each other assuming that **(H1.1)** holds.

Proposition 4.3. *Assume that* **(H1.1)** *is satisfied. Thus, almost surely,*

$$\left|\frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1\right| \leq \frac{b_n - a_n}{a_n},$$

with the convention that $0/0 = 1$.

Since KeRF estimates are kernel estimates of the form (4.1), Proposition 4.3 stresses that random forests are close to kernel estimates if the number of points in each cell is controlled. As highlighted by the following discussion, the assumptions of Proposition 4.3 are satisfied for some types of random forests.

Centred random forests of level k. For this model, whenever $\mathbf{X}$ is uniformly distributed over $[0,1]^d$, each cell has a Lebesgue-measure of 2^{-k}. Thus, fixing $\mathbf{x} \in [0,1]^d$, according to the law of the iterated logarithm, for all n large enough, almost surely,

$$\left|N_n(\mathbf{x}, \Theta) - \frac{n}{2^k}\right| \leq \frac{\sqrt{2n \log\log n}}{2}.$$

Consequently, **(H1.1)** is satisfied for $a_n = n2^{-k} - \sqrt{2n \log\log n}/2$ and $b_n = n2^{-k} + \sqrt{2n \log\log n}/2$. This yields, according to Proposition 4.3, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq \frac{\sqrt{2n \log\log n}}{n2^{-k} - \sqrt{2n \log\log n}/2}.$$

Thus, choosing for example $k = (\log_2 n)/3$, centred KeRF estimates are asymptotically equivalent to centred forest estimates as $n \to \infty$. The previous inequality can be extended to the case where $\mathbf{X}$ has a density f satisfying $c \leq f \leq C$, for some constants $0 < c < C < \infty$. In that case, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq \frac{\sqrt{2n \log\log n} + (C-c)n/2^k}{nc2^{-k} - \sqrt{2n \log\log n}/2}.$$

However, the right-hand term does not tend to zero as $n \to \infty$, meaning that the uniform assumption on $\mathbf{X}$ is crucial to prove the asymptotic equivalence of $m_{M,n}$ and $\widetilde{m}_{M,n}$ in the case of centred forests.

Breiman's forests. Each leaf in Breiman's trees contains a small number of points (typically between 1 and 5). Thus, if each cell contains exactly one point (default settings in classification problems), **(H1.1)** holds with $a_n = b_n = 1$. Thus, according to Proposition 4.3, almost surely,

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M).$$

More generally, if the number of observations in each cell varies between 1 and 5, one can set $a_n = 1$ and $b_n = 5$. Thus, still by Proposition 4.3, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq 4.$$

Median forests of level k. In this model, each cell of each tree is split at the empirical median of the observations belonging to the cell. The process is repeated until every cell is cut exactly k times (where $k \in \mathbb{N}$ is a parameter chosen by the user). Thus, each cell contains the same number of points ± 2 [see, e.g., Biau and Devroye, 2013, for details], and, according to Proposition 4.3, almost surely,

$$\left| \frac{m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)}{\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M)} - 1 \right| \leq \frac{2}{a_n}.$$

Consequently, if the level k of each tree is chosen such that $a_n \to \infty$ as $n \to \infty$, median KeRF estimates are equivalent to median forest estimates.

The following lemma extends Proposition 4.3 to infinite KeRF and forest estimates.

Proposition 4.4. *Assume that* **(H1.2)** *is satisfied. Thus, almost surely,*

$$|m_{\infty,n}(\mathbf{x}) - \widetilde{m}_{\infty,n}(\mathbf{x})| \leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + n\varepsilon_n \Big(\max_{1 \leq i \leq n} Y_i \Big).$$

Considering inequalities provided in Proposition 4.4, we see that infinite KeRF estimates are close to infinite random forest estimates if the number of observations in each cell is bounded (via a_n and b_n).

It is worth noticing that controlling the number of observations in each cell while obtaining a simple partition shape is difficult to achieve. On the one hand, if the tree construction depends on the training set, the algorithm can be stopped when each leaf contains exactly one point and thus KeRF estimate is equal to random forest estimate. However, in that case, the probability $K_n(\mathbf{x}, \mathbf{z})$ is very difficult to express since the geometry of each tree partitioning strongly depends on the training set. On the other hand, if the tree construction is independent of the training set, the probability $K_n(\mathbf{x}, \mathbf{z})$ can be made explicit in some cases, for example for centred forests (see Section 5). However, the number of points in each cell is difficult to control (every leaf cannot contain exactly one point with a non-adaptive cutting strategy) and thus KeRF estimate can be far away from random forest estimate. Consequently, one cannot deduce an explicit expression for random forest estimates from the explicit expression of KeRF estimates.

4.4 Two particular KeRF estimates

According to Proposition 4.2, infinite KeRF estimate $\widetilde{m}_{\infty,n}$ depends only on the connection function K_n via the following equation

$$\widetilde{m}_{\infty,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_n(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^{n} K_n(\mathbf{x}, \mathbf{X}_\ell)}. \tag{4.6}$$

To take one step further into the understanding of KeRF, we study in this section the connection function of two specific infinite random forests. We focus on infinite KeRF estimates for two reasons. Firstly, the expressions of infinite KeRF estimates are more amenable to mathematical analysis since they do not depend on the particular trees used to build the forest. Secondly, the prediction accuracy of infinite random forests is known to be better than that of finite random forests [see, e.g., Scornet, 2014]. Therefore infinite KeRF estimates are likely to be more accurate than finite KeRF estimates.

Practically, both infinite KeRF estimates and infinite random forest estimates can only be approximated by Monte Carlo simulations. Here, we show that centred KeRF estimates have an explicit expression, that is their connection function can be made explicit. Thus, infinite centred KeRF estimates and infinite uniform KeRF estimates (up to an approximation detailed below) can be directly computed using equation (4.6).

Centred KeRF As seen above, the construction of centred KeRF of level k is the same as for centred forests of level k except that predictions are made according to equation (4.2). Centred random forests are closely related to Breiman's forests in a linear regression framework. Indeed, in this context, splits that are performed at a low level of the trees are roughly located at the middle of each cell. In that case, Breiman's forests and centred forests are close to each other, which justifies the interest for these simplified models, and thus for centred KeRF.

In the sequel, the connection function of the centred random forest of level k is denoted by K_k^{cc}. This notation is justified by the fact that the construction of centred KeRF estimates depends only on the size of the training set through the choice of k.

Proposition 4.5. *Let $k \in \mathbb{N}$ and consider an infinite centred random forest of level k. Then, for all $\mathbf{x}, \mathbf{z} \in [0,1]^d$,*

$$K_k^{cc}(\mathbf{x}, \mathbf{z}) = \sum_{\substack{k_1,\dots,k_d \\ \sum_{\ell=1}^d k_\ell = k}} \frac{k!}{k_1! \dots k_d!} \left(\frac{1}{d}\right)^k \prod_{j=1}^d \mathbb{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} z_j \rceil}.$$

Note that ties are broken by imposing that cells are of the form $\prod_{i=1}^d A_i$ where the A_i are equal to $]a_i, b_i]$ or $[0, b_i]$, for all $0 < a_i < b_i \leq 1$. Figure 4.1 shows a graphical representation of the function f defined as

$$\begin{array}{rccc} f_k: & [0,1] \times [0,1] & \to & [0,1] \\ & \mathbf{z} = (z_1, z_2) & \mapsto & K_k^{cc}((\frac{1}{2}, \frac{1}{2}), \mathbf{z}). \end{array}$$

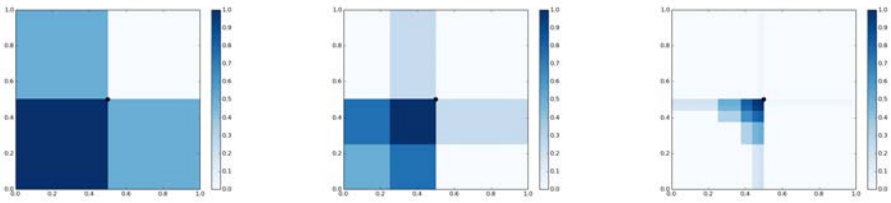

Figure 4.1: Representations of f_1, f_2 and f_5 in $[0,1]^2$

Denote by $\widetilde{m}_{\infty,n}^{cc}$ the infinite centred KeRF estimate, associated with the connection function K_k^{cc}, defined as

$$\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_k^{cc}(\mathbf{x}, \mathbf{X}_i)}{\sum_{\ell=1}^n K_k^{cc}(\mathbf{x}, \mathbf{X}_\ell)}.$$

To pursue the analysis of $\widetilde{m}_{\infty,n}^{cc}$, we will need the following assumption on the regression model.

(H4.2) (H2) *One has*

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a centred Gaussian noise, independent of $\mathbf{X}$, with finite variance $\sigma^2 < \infty$. Moreover, $\mathbf{X}$ is uniformly distributed on $[0,1]^d$ and m is Lipschitz.

Our theorem states that infinite centred KeRF estimates are consistent whenever **(H2)** holds. Moreover, it provides an upper bound on the rate of consistency of centred KeRF.

Theorem 4.1. *Assume that* **(H2)** *is satisfied. Then, providing $k \to \infty$ and $n/2^k \to \infty$, there exists a constant $C_1 > 0$ such that, for all $n > 1$, and for all $\mathbf{x} \in [0,1]^d$,*

$$\mathbb{E}\left[\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq C_1 n^{-1/(3+d\log 2)} (\log n)^2.$$

Observe that centred KeRF estimates fail to reach minimax rate of consistency $n^{-2/(d+2)}$ over the class of Lipschitz functions. A similar upper bound on the rate of consistency $n^{-3/4d\log 2+3}$ of centred random forests was obtained by Biau [2012]. It is worth noticing that, for all $d \geq 9$, the upper bound on the rate of centred KeRF is sharper than that of centred random forests. This theoretical result supports the fact that KeRF procedure has a better performance compared to centred random forests. This will be supported by simulations in Section 5 (see Figure 4.5)

Uniform KeRF Recall that the infinite uniform KeRF estimates of level k are the same as infinite uniform forest of level k except that predictions are computed according to equation (4.2). Uniform random forests, first studied by Biau et al. [2008], remain under active investigation. They are a nice modelling of Breiman forests, since with no a priori on the split location, we can consider that splits are drawn uniformly on the cell edges. Other related versions of these forests have been thoroughly investigated by Arlot and Genuer [2014] who compare the bias of a single tree to that of the whole forest.

As for the connection function of centred random forests, we use the notational convention K_k^{uf} to denote the connection function of uniform random forests of level k.

Proposition 4.6. *Let $k \in \mathbb{N}$ and consider an infinite uniform random forest of level k. Then, for all $\mathbf{x} \in [0,1]^d$,*

$$K_k^{uf}(\mathbf{0},\mathbf{x}) = \sum_{\substack{k_1,\dots,k_d \\ \sum_{\ell=1}^d k_\ell = k}} \frac{k!}{k_1!\dots k_d!}\left(\frac{1}{d}\right)^k \prod_{m=1}^{d}\left(1 - x_m \sum_{j=0}^{k_m-1}\frac{(-\ln x_m)^j}{j!}\right),$$

with the convention $\sum_{j=0}^{-1}\frac{(-\ln x_m)^j}{j!} = 0$.

Proposition 4.6 gives the explicit expression of $K_k^{uf}(\mathbf{0},\mathbf{x})$. Figure 4.2 shows a representation of the functions f_1, f_2 and f_5 defined as

$$\begin{array}{rccl} f_k: & [0,1]\times[0,1] & \to & [0,1] \\ & \mathbf{z}=(z_1,z_2) & \mapsto & K_k^{uf}(\mathbf{0},|\mathbf{z}-(\frac{1}{2},\frac{1}{2})|), \end{array}$$

where $|\mathbf{z}-\mathbf{x}| = (|z_1-x_1|,\dots,|z_d-x_d|)$.

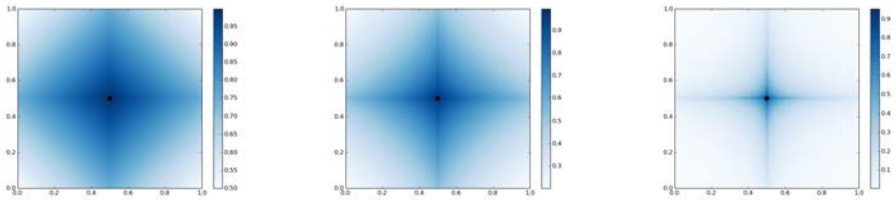

Figure 4.2: Representations of f_1, f_2 and f_5 in dimension two

Unfortunately, the general expression of the connection function $K_k^{uf}(\mathbf{x},\mathbf{z})$ is difficult to obtain. Indeed, for $d = 1$, cuts are performed along a single axis, but the probability of connection

between two points x and z does not depend only upon the distance $|z-x|$ but rather on the positions x and z, as stressed in the following Lemma.

Lemma 4.1. *Let $x, z \in [0,1]$. Then,*

$$K_1^{uf}(x,z) = 1 - |z-x|,$$
$$K_2^{uf}(x,z) = 1 - |z-x| + |z-x| \log\left(\frac{z}{1-x}\right).$$

A natural way to deal with this difficulty is to replace the connection function K_k^{uf} by the function $(\mathbf{x},\mathbf{z}) \to K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)$. Indeed, this is a simple manner to build an invariant-by-translation version of the uniform kernel K_k^{uf}. The extensive simulations in Section 5 support the fact that estimates of the form (4.6) built with these two kernels have similar prediction accuracy. As for infinite centred KeRF estimates, we denote by $\widetilde{m}_{\infty,n}^{uf}$ the infinite uniform KeRF estimates but built with the invariant-by-translation version of K_k^{uf}, namely

$$\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\sum_{\ell=1}^n K_k^{uf}(\mathbf{0}, |\mathbf{X}_\ell - \mathbf{x}|)}.$$

Our last theorem states the consistency of infinite uniform KeRF estimates along with an upper bound on their rate of consistency.

Theorem 4.2. *Assume that* **(H2)** *is satisfied. Then, providing $k \to \infty$ and $n/2^k \to \infty$, there exists a constant $C_1 > 0$ such that, for all $n > 1$ and for all $\mathbf{x} \in [0,1]^d$,*

$$\mathbb{E}\left[\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq C_1 n^{-2/(6+3d\log 2)} (\log n)^2.$$

As for centred KeRF estimates, the rate of consistency does not reach the minimax rate on the class of Lipschitz functions, and is actually worse than that of centred KeRF estimates, whatever the dimension d is. Besides, centred KeRF estimates have better performance than uniform KeRF estimates and this will be highlighted by simulations (Section 5).

Although centred and uniform KeRF estimates are kernel estimates of the form (4.1), the usual tools used to prove consistency and to find rate of consistency of kernel methods cannot be applied here [see, e.g., Chapter 5 in Györfi et al., 2002]. Indeed, the support of $\mathbf{z} \mapsto K_k^{cc}(\mathbf{x},\mathbf{z})$ and that of $\mathbf{z} \mapsto K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)$ cannot be contained in a ball centred on $\mathbf{x}$, whose diameter tends to zero (see Figure 4.1 and 4.2). The proof of Theorem 4.1 and 4.2 are then based on the previous work of Greblicki et al. [1984] who proved the consistency of kernels with unbounded support. In particular, we use their bias/variance decomposition of kernel estimates to exhibit upper bounds on the rate of consistency.

4.5 Experiments

Practically speaking, Breiman's random forests are among the most widely used forest algorithms. Thus a natural question is to know whether Breiman KeRF compare favourably to

Breiman's forests. In fact, as seen above, the two algorithms coincide whenever Breiman's forests are fully grown. But this is not always the case since by default, each cell of Breiman's forests contain between 1 and 5 observations.

We start this section by comparing Breiman KeRF and Breiman's forest estimates for various regression models described below. Some of these models are toy models (**Model 1, 5-8**). **Model 2** can be found in van der Laan et al. [2007] and **Models 3-4** are presented in Meier et al. [2009]. For all regression frameworks, we consider covariates $\mathbf{X} = (X_1, \ldots, X_d)$ that are uniformly distributed over $[0,1]^d$. We also let $\widetilde{X}_i = 2(X_i - 0.5)$ for $1 \leq i \leq d$.

Model 1: $n = 800, d = 50, Y = \widetilde{X}_1^2 + \exp(-\widetilde{X}_2^2)$

Model 2: $n = 600, d = 100, Y = \widetilde{X}_1\widetilde{X}_2 + \widetilde{X}_3^2 - \widetilde{X}_4\widetilde{X}_7 + \widetilde{X}_8\widetilde{X}_{10} - \widetilde{X}_6^2 + \mathcal{N}(0, 0.5)$

Model 3: $n = 600, d = 100, Y = -\sin(2\widetilde{X}_1) + \widetilde{X}_2^2 + \widetilde{X}_3 - \exp(-\widetilde{X}_4) + \mathcal{N}(0, 0.5)$

Model 4: $n = 600, d = 100, Y = \widetilde{X}_1 + (2\widetilde{X}_2 - 1)^2 + \sin(2\pi\widetilde{X}_3)/(2 - \sin(2\pi\widetilde{X}_3)) + \sin(2\pi\widetilde{X}_4) + 2\cos(2\pi\widetilde{X}_4) + 3\sin^2(2\pi\widetilde{X}_4) + 4\cos^2(2\pi\widetilde{X}_4) + \mathcal{N}(0, 0.5)$

Model 5: $n = 700, d = 20, Y = \mathbb{1}_{\widetilde{X}_1 > 0} + \widetilde{X}_2^3 + \mathbb{1}_{\widetilde{X}_4 + \widetilde{X}_6 - \widetilde{X}_8 - \widetilde{X}_9 > 1 + \widetilde{X}_{10}} + \exp(-\widetilde{X}_2^2) + \mathcal{N}(0, 0.5)$

Model 6: $n = 500, d = 30, Y = \sum_{k=1}^{10} \mathbb{1}_{\widetilde{X}_k^3 < 0} - \mathbb{1}_{\mathcal{N}(0,1) > 1.25}$

Model 7: $n = 600, d = 300, Y = \widetilde{X}_1^2 + \widetilde{X}_2^2\widetilde{X}_3 \exp(-|\widetilde{X}_4|) + \widetilde{X}_6 - \widetilde{X}_8 + \mathcal{N}(0, 0.5)$

Model 8: $n = 500, d = 1000, Y = \widetilde{X}_1 + 3\widetilde{X}_3^2 - 2\exp(-\widetilde{X}_5) + \widetilde{X}_6$

All numerical implementations have been performed using the free Python software, available online at `https://www.python.org/`. For each experiment, the data set is divided into a training set (80% of the data set) and a test set (the remaining 20%). Then, the empirical risk ($\mathbb{L}^2$ error) is evaluated on the test set.

To start with, Figure 4.3 depicts the empirical risk of Breiman's forests and Breiman KeRF estimates for two regression models (the conclusions are similar for the remaining regression models). Default settings were used for Breiman's forests (`minsamplessplit` $= 2$, `maxfeatures` $= 0.333$) and for Breiman KeRF, except that we did not bootstrap the data set. Figure 4.3 puts in evidence that Breiman KeRF estimates behave similarly (in terms of empirical risk) to Breiman forest estimates. It is also interesting to note that bootstrapping the data set does not change the performance of the two algorithms.

Figure 4.4 (resp. Figure 4.5) shows the risk of uniform (resp. centred) KeRF estimates compared to the risk of uniform (resp. centred) forest estimates (only two models shown). In these two experiments, uniform and centred forests and their KeRF counterparts have been grown in such a way that each tree is a complete binary tree of level $k = \lfloor \log_2 n \rfloor$. Thus, in that case, each cell contains on average $n/2^k \simeq 1$ observation. Once again, the main message of Figure 4.4 is that the uniform KeRF accuracy is close to the uniform forest accuracy.

On the other hand, it turns out that the performance of centred KeRF and centred forests are not similar (Figure 4.5). In fact, centred KeRF estimates are either comparable to centred forest estimates (as, for example, in **Model 2**), or have a better accuracy (as, for example, in

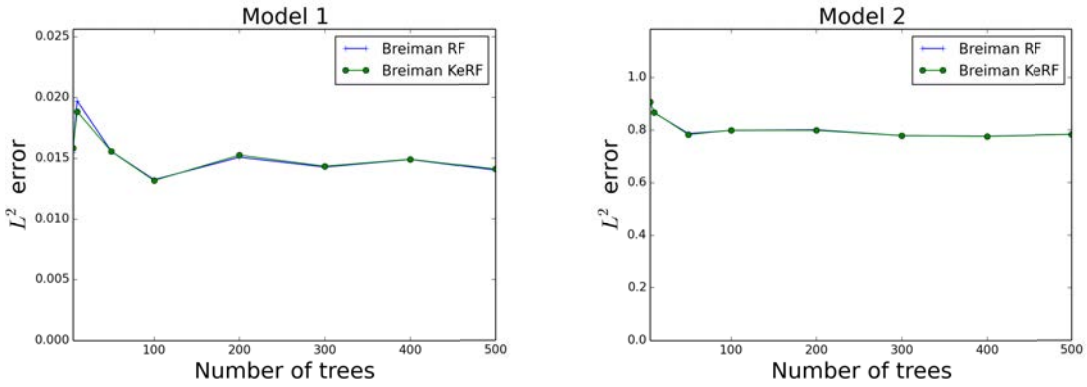

Figure 4.3: Empirical risks of Breiman KeRF estimates and Breiman forest estimates.

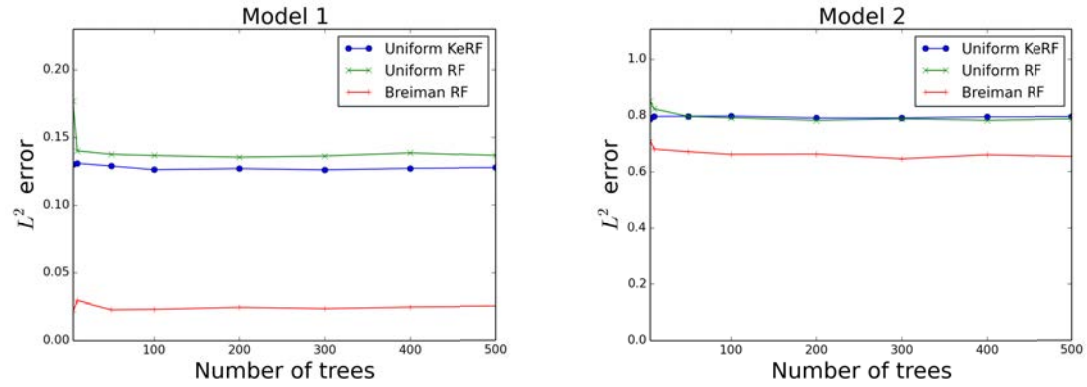

Figure 4.4: Empirical risks of uniform KeRF and uniform forest.

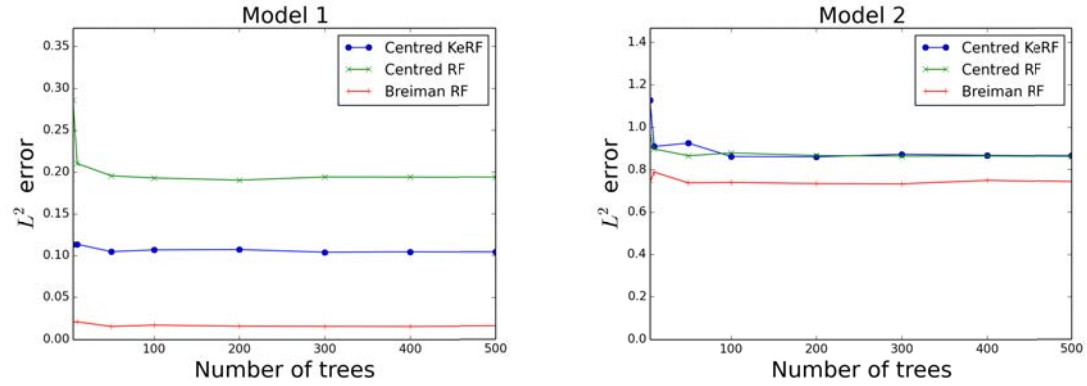

Figure 4.5: Empirical risks of centred KeRF and centred forest.

Model 1). A possible explanation for this phenomenon is that centred forests are non-adaptive in the sense that their construction does not depend on the data set. Therefore, each tree is likely to contain cells with unbalanced number of data points, which can result in random forest misestimation. This undesirable effect vanishes using KeRF methods since they assign the same weights to each observation.

The same series of experiments were conducted, but using bootstrap for computing both KeRF and random forest estimates. The general finding is that the results are similar—Figure 4.6 and 4.7 depict the accuracy of corresponding algorithms for a selected choice of regression frameworks.

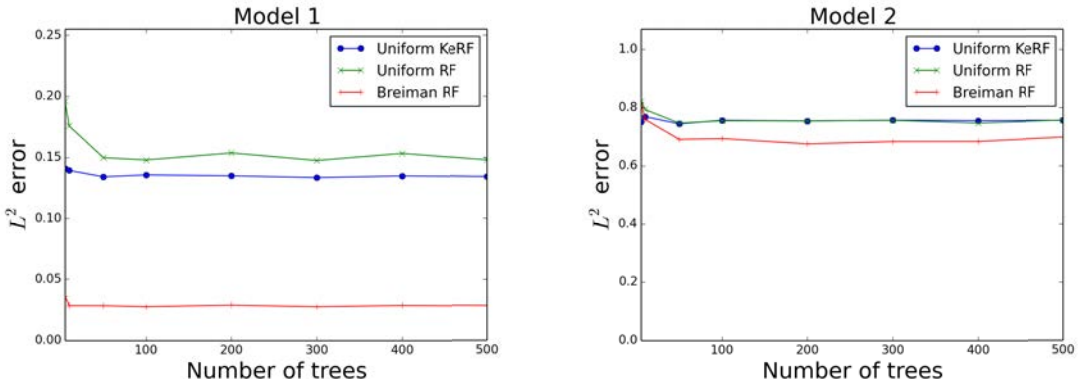

Figure 4.6: Empirical risks of uniform KeRF and uniform forest (with bootstrap).

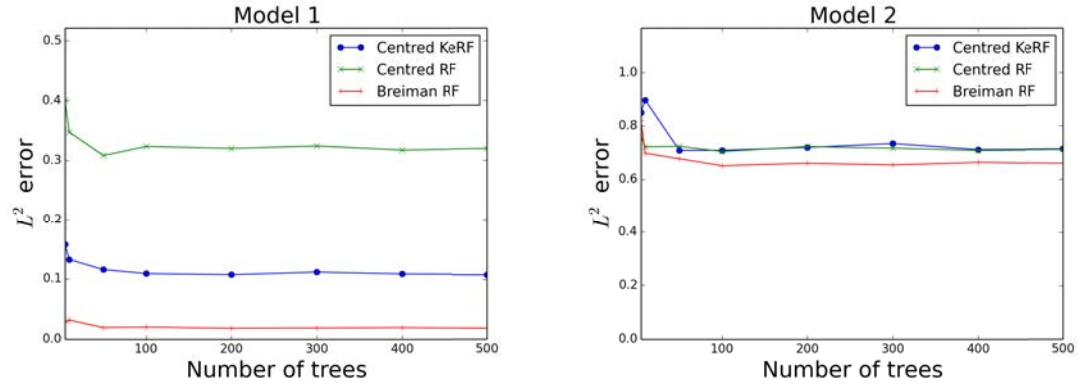

Figure 4.7: Empirical risks of centred KeRF and centred forests (with bootstrap).

An important aspect of infinite centred and uniform KeRF is that they can be explicitly computed (see Proposition 4.5 and 4.6). Thus, we have plotted in Figure 4.8 the empirical risk of both finite and infinite centred KeRF estimates for some examples (for $n = 100$ and $d = 10$). We clearly see in this figure that the accuracy of finite centred KeRF tends to the accuracy of infinite centred KeRF as M tends to infinity. This corroborates Proposition 4.2.

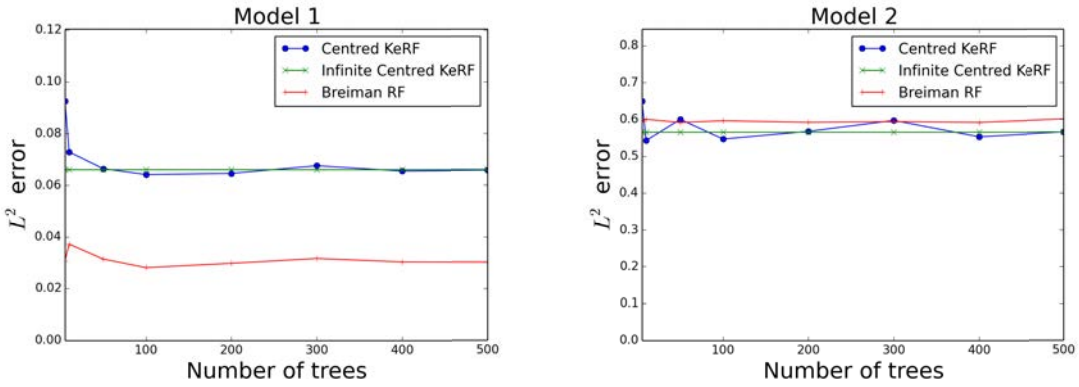

Figure 4.8: Risks of finite and infinite centred KeRF.

The same comments hold for uniform KeRF (see Figure 4.9). Note however that, in that case, the proximity between finite uniform KeRF and infinite uniform KeRF estimate strengthens the approximation that has been made on infinite uniform KeRF in Section 4.

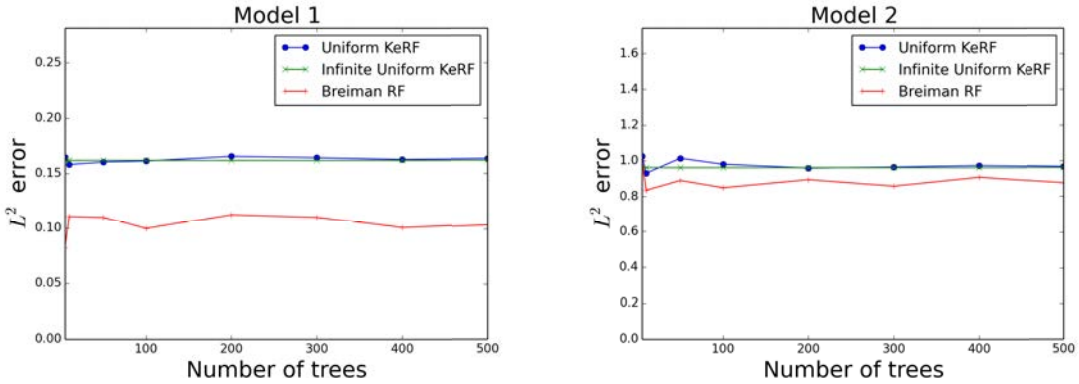

Figure 4.9: Risks of finite and infinite uniform KeRF.

The computation time for finite KeRF estimate is very acceptable for finite KeRF and similar to that of random forest (Figure 4.3-4.5). However, the story is different for infinite KeRF estimates. In fact, KeRF estimates can only be evaluated for low dimensional data sets and small sample sizes. To see this, just note that the explicit formulation of KeRF involves a multinomial distribution (Proposition 4.5 and 4.6). Each evaluation of the multinomial creates computational burden when the dimensions (d and n) of the problems increases. For example, in Figure 4.8 and 4.9, the computation time needed to compute infinite KeRF estimates ranges between thirty minutes to 3 hours. As a matter of fact, infinite KeRF methods should be seen as theoretical tools rather than a practical substitute for random forests.

4.6 Proofs

Proof of Proposition 4.1. By definition,

$$\begin{aligned}\widetilde{m}_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) &= \frac{1}{\sum_{j=1}^{M} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)} \\ &= \frac{M}{\sum_{j=1}^{M} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)}} \sum_{i=1}^{n} Y_i K_{M,n}(\mathbf{x}, \mathbf{X}_i).\end{aligned}$$

Finally, observe that

$$\frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta_j)} = \sum_{i=1}^{n} K_{M,n}(\mathbf{x}, \mathbf{X}_i),$$

which concludes the proof. □

Proof of Proposition 4.2. We prove the result for $d = 2$. The other cases can be treated similarly. For the moment, we assume the random forest to be continuous. Recall that, for all $\mathbf{x}, \mathbf{z} \in [0,1]^2$, and for all $M \in \mathbb{N}$,

$$K_{M,n}(\mathbf{x}, \mathbf{z}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}_{\mathbf{z} \in A_n(\mathbf{x}, \Theta_j)}.$$

According to the strong law of large numbers, almost surely, for all $\mathbf{x}, \mathbf{z} \in \mathbb{Q}^2 \cap [0,1]^2$

$$\lim_{M \to \infty} K_{M,n}(\mathbf{x}, \mathbf{z}) = K_n(\mathbf{x}, \mathbf{z}).$$

Set $\varepsilon > 0$ and $\mathbf{x}, \mathbf{z} \in [0,1]^2$ where $\mathbf{x} = (x^{(1)}, x^{(2)})$ and $\mathbf{z} = (z^{(1)}, z^{(2)})$. Assume, without loss of generality, that $x^{(1)} < z^{(1)}$ and $x^{(2)} < z^{(2)}$. Let

$$\begin{aligned}A_{\mathbf{x}} &= \{\mathbf{u} \in [0,1]^2, u^{(1)} \leq x^{(1)} \text{ and } u^{(2)} \leq x^{(2)}\}, \\ \text{and } A_{\mathbf{z}} &= \{\mathbf{u} \in [0,1]^2, u^{(1)} \geq z^{(1)} \text{ and } u^{(2)} \geq z^{(2)}\}.\end{aligned}$$

Choose $\mathbf{x}_1 \in A_{\mathbf{x}} \cap \mathbb{Q}^2$ (resp. $\mathbf{z}_2 \in A_{\mathbf{z}} \cap \mathbb{Q}^2$) and take $\mathbf{x}_2 \in [0,1]^2 \cap \mathbb{Q}^2$ (resp. $\mathbf{z}_1 \in [0,1]^2 \cap \mathbb{Q}^2$) such that $x_1^{(1)} \leq x^{(1)} \leq x_2^{(1)}$ and $x_1^{(2)} \leq x^{(2)} \leq x_2^{(2)}$ (resp. $z_1^{(1)} \leq z^{(1)} \leq z_2^{(1)}$ and $z_1^{(2)} \leq z^{(2)} \leq z_2^{(2)}$, see Figure 4.10).

Observe that, because of the continuity of K_n, one can choose $\mathbf{x}_1, \mathbf{x}_2$ close enough to $\mathbf{x}$ and $\mathbf{z}_2, \mathbf{z}_1$ close enough to $\mathbf{z}$ such that

$$\begin{aligned}|K_n(\mathbf{x}_2, \mathbf{x}_1) - 1| &\leq \varepsilon, \\ |K_n(\mathbf{z}_1, \mathbf{z}_2) - 1| &\leq \varepsilon, \\ \text{and } |K_n(\mathbf{x}_1, \mathbf{z}_2) - K_n(\mathbf{x}, \mathbf{z})| &\leq \varepsilon.\end{aligned}$$

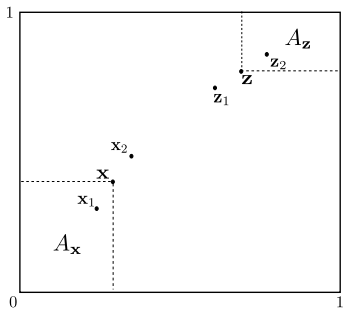

Figure 4.10: Respective positions of $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2$

Bounding the difference between $K_{M,n}$ and K_n, we have

$$\begin{aligned}|K_{M,n}(\mathbf{x},\mathbf{z}) - K_n(\mathbf{x},\mathbf{z})| \leq\, & |K_{M,n}(\mathbf{x},\mathbf{z}) - K_{M,n}(\mathbf{x}_1,\mathbf{z}_2)| \\ & + |K_{M,n}(\mathbf{x}_1,\mathbf{z}_2) - K_n(\mathbf{x}_1,\mathbf{z}_2)| \\ & + |K_n(\mathbf{x}_1,\mathbf{z}_2) - K_n(\mathbf{x},\mathbf{z})|\,. \end{aligned} \tag{4.7}$$

To simplify notation, we let $\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}$ be the event where $\mathbf{x}$ and $\mathbf{z}$ are in the same cell in the tree built with randomness Θ_j and dataset $\mathcal{D}_n$. We also let $\mathbf{x} \overset{\Theta_m}{\nleftrightarrow} \mathbf{z}$ be the complement event of $\mathbf{x} \overset{\Theta_m}{\leftrightarrow} \mathbf{z}$. Accordingly, the first term on the right side in equation (4.7) is bounded above by

$$\begin{aligned}|K_{M,n}(\mathbf{x},\mathbf{z}) - K_{M,n}(\mathbf{x}_1,\mathbf{z}_2)| &\leq \frac{1}{M}\sum_{m=1}^{M}\left|\mathbb{1}_{\mathbf{x}\overset{\Theta_m}{\leftrightarrow}\mathbf{z}} - \mathbb{1}_{\mathbf{x}_1\overset{\Theta_m}{\leftrightarrow}\mathbf{z}_2}\right| \\ &\leq \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}_{\mathbf{x}_1\overset{\Theta_m}{\nleftrightarrow}\mathbf{x}} + \mathbb{1}_{\mathbf{z}_2\overset{\Theta_m}{\nleftrightarrow}\mathbf{z}} \\ &\quad \text{(given the positions of } \mathbf{x}, \mathbf{x}_1, \mathbf{z}, \mathbf{z}_2) \\ &\leq \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}_{\mathbf{x}_1\overset{\Theta_m}{\nleftrightarrow}\mathbf{x}_2} + \frac{1}{M}\sum_{m=1}^{M}\mathbb{1}_{\mathbf{z}_2\overset{\Theta_m}{\nleftrightarrow}\mathbf{z}_1}, \end{aligned} \tag{4.8}$$

given the respective positions of $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2$. But, since $\mathbf{x}_2, \mathbf{z}_1, \mathbf{x}_1, \mathbf{z}_2 \in \mathbb{Q}^2 \cap [0,1]^2$, we deduce from inequation (4.8) that, for all M large enough,

$$|K_{M,n}(\mathbf{x},\mathbf{z}) - K_{M,n}(\mathbf{x}_1,\mathbf{z}_2)| \leq 1 - K_n(\mathbf{x}_2,\mathbf{x}_1) + 1 - K_n(\mathbf{z}_1,\mathbf{z}_2) + 2\varepsilon.$$

Combining the last inequality with equation (4.7), we obtain, for all M large enough,

$$\begin{aligned}|K_{M,n}(\mathbf{x},\mathbf{z}) - K_n(\mathbf{x},\mathbf{z})| \leq\, & 1 - K_n(\mathbf{x}_2,\mathbf{x}_1) + 1 - K_n(\mathbf{z}_1,\mathbf{z}_2) \\ & + |K_{M,n}(\mathbf{x}_1,\mathbf{z}_2) - K_n(\mathbf{x}_1,\mathbf{z}_2)| \\ & + |K_n(\mathbf{x}_1,\mathbf{z}_2) - K_n(\mathbf{x},\mathbf{z})| + 2\varepsilon \\ \leq\, & 6\varepsilon. \end{aligned}$$

Consequently, for any continuous random forest, almost surely, for all $\mathbf{x}, \mathbf{z} \in [0,1]^2$,

$$\lim_{M\to\infty} K_{M,n}(\mathbf{x},\mathbf{z}) = K_n(\mathbf{x},\mathbf{z}).$$

The proof can be easily adapted to the case of discrete random forests. Thus, this complete the first part of the proof. Next, observe that

$$\lim_{M\to\infty} \frac{\sum_{i=1}^n Y_i K_{M,n}(\mathbf{x},\mathbf{X}_i)}{\sum_{j=1}^n K_{M,n}(\mathbf{x},\mathbf{X}_j)} = \frac{\sum_{i=1}^n Y_i K_n(\mathbf{x},\mathbf{X}_i)}{\sum_{j=1}^n K_n(\mathbf{x},\mathbf{X}_j)},$$

for all $\mathbf{x}$ satisfying $\sum_{j=1}^n K_n(\mathbf{x},\mathbf{X}_j) \neq 0$. Thus, almost surely for those $\mathbf{x}$,

$$\lim_{M\to\infty} \widetilde{m}_{M,n}(\mathbf{x}) = \widetilde{m}_{\infty,n}(\mathbf{x}). \tag{4.9}$$

Now, if there exists any $\mathbf{x}$ such that $\sum_{j=1}^n K_n(\mathbf{x},\mathbf{X}_j) = 0$, then $\mathbf{x}$ is not connected with any data points in any tree of the forest. In that case, $\sum_{j=1}^n K_{M,n}(\mathbf{x},\mathbf{X}_j) = 0$ and, by convention, $\widetilde{m}_{\infty,n}(\mathbf{x}) = \widetilde{m}_{M,n}(\mathbf{x}) = 0$. Finally, formula (4.9) holds for all $\mathbf{x} \in [0,1]^2$. □

Proof of Proposition 4.3. Fix $\mathbf{x} \in [0,1]^d$ and assume that, a.s., $Y \geq 0$. By assumption **(H1.1)**, there exist sequences $(a_n), (b_n)$ such that, almost surely,

$$a_n \leq N_n(\mathbf{x},\Theta) \leq b_n.$$

To simplify notation, we let $\bar{N}_{M,n}(\mathbf{x},\Theta) = \frac{1}{M}\sum_{j=1}^M N_n(\mathbf{x},\Theta_j)$. Thus, almost surely,

$$\begin{aligned}
|m_{M,n}(\mathbf{x}) - \widetilde{m}_{M,n}(\mathbf{x})| &= \left| \sum_{i=1}^n Y_i \left(\frac{1}{M}\sum_{m=1}^M \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta_m)}}{N_n(\mathbf{x},\Theta_m)} \right) \right. \\
&\qquad \left. - \sum_{i=1}^n Y_i \left(\frac{1}{M}\sum_{m=1}^M \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta_m)}}{\bar{N}_{M,n}(\mathbf{x})} \right) \right| \\
&\leq \frac{1}{M}\sum_{i=1}^n Y_i \sum_{m=1}^M \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta_m)}}{\bar{N}_{M,n}(\mathbf{x})} \times \left| \frac{\bar{N}_{M,n}(\mathbf{x})}{N_n(\mathbf{x},\Theta_m)} - 1 \right| \\
&\leq \frac{b_n - a_n}{a_n} \widetilde{m}_{M,n}(\mathbf{x}).
\end{aligned}$$

□

Proof of Proposition 4.4. Fix $\mathbf{x} \in [0,1]^d$ and assume that, almost surely, $Y \geq 0$. By assumption **(H1.2)**, there exist sequences (a_n), (b_n), (ε_n) such that, letting A be the event where

$$a_n \leq N_n(\mathbf{x},\Theta) \leq b_n,$$

we have, almost surely,

$$\mathbb{P}_\Theta[A] \geq 1 - \varepsilon_n \quad \text{and} \quad 1 \leq a_n \leq \mathbb{E}_\Theta\left[N_n(\mathbf{x},\Theta)\right] \leq b_n.$$

Therefore, a.s.,

$$\begin{aligned}
&|m_{\infty,n}(\mathbf{x}) - \widetilde{m}_{\infty,n}(\mathbf{x})| \\
&= \Big| \sum_{i=1}^{n} Y_i \mathbb{E}_\Theta \left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{N_n(\mathbf{x},\Theta)} \right] - \sum_{i=1}^{n} Y_i \mathbb{E}_\Theta \left[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{\mathbb{E}_\Theta \left[N_n(\mathbf{x},\Theta) \right]} \right] \Big| \\
&= \Big| \sum_{i=1}^{n} Y_i \mathbb{E}_\Theta \Big[\frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x},\Theta)}}{\mathbb{E}_\Theta \left[N_n(\mathbf{x},\Theta) \right]} \Big(\frac{\mathbb{E}_\Theta \left[N_n(\mathbf{x},\Theta) \right]}{N_n(\mathbf{x},\Theta)} - 1 \Big) (\mathbb{1}_A + \mathbb{1}_{A^c}) \Big] \Big| \\
&\leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + \Big(\max_{1 \leq i \leq n} Y_i \Big) \mathbb{E}_\Theta \Big[\Big| 1 - \frac{N_n(\mathbf{x},\Theta)}{\mathbb{E}_\Theta \left[N_n(\mathbf{x},\Theta) \right]} \Big| \mathbb{1}_{A^c} \Big] \\
&\leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + n \Big(\max_{1 \leq i \leq n} Y_i \Big) \mathbb{P}[A^c].
\end{aligned}$$

Consequently, almost surely,

$$|m_{\infty,n}(\mathbf{x}) - \widetilde{m}_{\infty,n}(\mathbf{x})| \leq \frac{b_n - a_n}{a_n} \widetilde{m}_{\infty,n}(\mathbf{x}) + n\varepsilon_n \Big(\max_{1 \leq i \leq n} Y_i \Big).$$

□

Proof of Proposition 4.5. Assume for the moment that $d = 1$. Take $x, z \in [0,1]$ and assume, without loss of generality, that $x \leq z$. Then the probability that x and z be in the same cell, after k cuts, is equal to

$$\mathbb{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} z_j \rceil}.$$

To prove the result in the multivariate case, take $\mathbf{x}, \mathbf{z} \in [0,1]^d$. Since cuts are independent, the probability that $\mathbf{x}$ and $\mathbf{z}$ are in the same cell after k cuts is given by the following multinomial

$$K_k^{cc}(\mathbf{x},\mathbf{z}) = \sum_{\substack{k_1,\ldots,k_d \\ \sum_{\ell=1}^d k_\ell = k}} \frac{k!}{k_1! \ldots k_d!} \prod_{j=1}^{d} \left(\frac{1}{d} \right)^{k_j} \mathbb{1}_{\lceil 2^{k_j} x_j \rceil = \lceil 2^{k_j} z_j \rceil}.$$

□

To prove Theorem 4.1, we need to control the bias of the centred KeRF estimate, which is done in Theorem 4.3.

Theorem 4.3. *Let f be a L-Lipschitz function. Then, for all k,*

$$\sup_{\mathbf{x} \in [0,1]^d} \left| \frac{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) f(\mathbf{z}) \mathrm{d}z_1 \ldots \mathrm{d}z_d}{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) \mathrm{d}z_1 \ldots \mathrm{d}z_d} - f(\mathbf{x}) \right| \leq Ld \left(1 - \frac{1}{2d} \right)^k.$$

Proof of Theorem 4.3. Let $\mathbf{x} \in [0,1]^d$ and $k \in \mathbb{N}$. Take f a L-Lipschitz function. In the rest of the proof, for clarity reasons, we use the notation $\mathrm{d}\mathbf{z}$ instead of $\mathrm{d}z_1 \ldots \mathrm{d}z_d$. Thus,

$$\left| \frac{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) f(\mathbf{z}) \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) \mathrm{d}\mathbf{z}} - f(\mathbf{x}) \right| \leq \frac{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) |f(\mathbf{z}) - f(\mathbf{x})| \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) \mathrm{d}\mathbf{z}}.$$

Note that,

$$\begin{aligned}
&\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z})|f(\mathbf{z})-f(\mathbf{x})|\mathrm{d}\mathbf{z} \\
&\quad\leq L\sum_{\ell=1}^{d}\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z})|z_\ell - x_\ell|\mathrm{d}\mathbf{z} \\
&\quad\leq L\sum_{\ell=1}^{d}\int_{[0,1]^d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j=k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{d}\right)^k \prod_{m\neq\ell}\int_0^1 K_{k_m}^{cc}(x_m,z_m)\mathrm{d}z_m \\
&\qquad\qquad \times\int_0^1 K_{k_\ell}^{cc}(x_\ell,z_\ell)|z_\ell-x_\ell|\mathrm{d}z_\ell. \qquad (4.10)
\end{aligned}$$

The last integral is upper bounded by

$$\begin{aligned}
\int_{[0,1]} K_{k_\ell}^{cc}(x_\ell,z_\ell)|x_\ell-z_\ell|\mathrm{d}z_\ell &= \int_{[0,1]} \mathbb{1}_{\lceil 2^{k_\ell}x_\ell\rceil=\lceil 2^{k_\ell}z_\ell\rceil}|x_\ell-z_\ell|\mathrm{d}z_\ell \\
&\leq \left(\frac{1}{2}\right)^{k_\ell}\int_{[0,1]} \mathbb{1}_{\lceil 2^{k_\ell}x_\ell\rceil=\lceil 2^{k_\ell}z_\ell\rceil}\mathrm{d}z_\ell \\
&\leq \left(\frac{1}{2}\right)^{k_\ell}\int_{[0,1]} K_{k_\ell}^{cc}(x_\ell,z_\ell)\mathrm{d}z_\ell. \qquad (4.11)
\end{aligned}$$

Therefore, combining inequalities (4.10) and (4.11), we obtain,

$$\begin{aligned}
&\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z})|f(\mathbf{z})-f(\mathbf{x})|\mathrm{d}\mathbf{z} \\
&\quad\leq L\sum_{\ell=1}^{d}\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j=k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{2}\right)^{k_\ell}\left(\frac{1}{d}\right)^k \prod_{m=1}^{d}\int_0^1 K_{k_m}^{cc}(x_m,z_m)\mathrm{d}z_m \\
&\quad\leq L\left(\frac{1}{d}\right)^k\sum_{\ell=1}^{d}\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j=k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{2}\right)^{k_\ell+k}, \qquad (4.12)
\end{aligned}$$

since, simple calculations show that, for all $x_m \in [0,1]$ and for all $k_m \in \mathbb{N}$,

$$\int_0^1 K_{k_m}^{cc}(x_m,z_m)\mathrm{d}z_m = \int_{[0,1]} \mathbb{1}_{\lceil 2^{k_m}x_m\rceil=\lceil 2^{k_m}z_m\rceil}dz_m = \left(\frac{1}{2}\right)^{k_m}. \qquad (4.13)$$

Consequently, we get from inequality (4.12) that

$$\frac{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z})|f(\mathbf{z})-f(\mathbf{x})|\mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{z}} \leq L\left(\frac{1}{d}\right)^k\sum_{\ell=1}^{d}\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j=k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{2}\right)^{k_\ell}.$$

Taking the first term of the sum, we obtain

$$\left(\frac{1}{d}\right)^k \sum_{\substack{k_1,\dots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\dots k_d!}\left(\frac{1}{2}\right)^{k_1} = \sum_{k_1=0}^{k}\left(\frac{1}{2d}\right)^{k_1}\left(1-\frac{1}{d}\right)^{k-k_1}\frac{k!}{k_1!(k-k_1)!}$$

$$\leq \left(1-\frac{1}{2d}\right)^k.$$

Finally,

$$\frac{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z})|f(\mathbf{z})-f(\mathbf{x})|\mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{z}} \leq Ld\left(1-\frac{1}{2d}\right)^k.$$

□

Proof of Theorem 4.1. Let $\mathbf{x}\in[0,1]^d$, $\|m\|_\infty = \sup_{\mathbf{x}\in[0,1]^d}|m(\mathbf{x})|$ and recall that

$$\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_k^{cc}(\mathbf{x},\mathbf{X}_i)}{\sum_{i=1}^n K_k^{cc}(\mathbf{x},\mathbf{X}_i)}.$$

Thus, letting

$$A_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n\left(\frac{Y_i K_k^{cc}(\mathbf{x},\mathbf{X}_i)}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]} - \frac{\mathbb{E}\left[Y K_k^{cc}(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]}\right),$$

$$B_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n\left(\frac{K_k^{cc}(\mathbf{x},\mathbf{X}_i)}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]} - 1\right),$$

$$\text{and } M_n(\mathbf{x}) = \frac{\mathbb{E}\left[Y K_k^{cc}(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]},$$

the estimate $\widetilde{m}_{\infty,n}^{cc}(\mathbf{x})$ can be rewritten as

$$\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) = \frac{M_n(\mathbf{x})+A_n(\mathbf{x})}{1+B_n(\mathbf{x})},$$

which leads to

$$\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x}) = \frac{M_n(\mathbf{x}) - m(\mathbf{x}) + A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})}{1+B_n(\mathbf{x})}.$$

According to Theorem 4.3, we have

$$\begin{aligned}|M_n(\mathbf{x}) - m(\mathbf{x})| &= \left|\frac{\mathbb{E}\left[m(\mathbf{X})K_k^{cc}(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]} + \frac{\mathbb{E}\left[\varepsilon K_k^{cc}(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]} - m(\mathbf{x})\right| \\ &\leq \left|\frac{\mathbb{E}\left[m(\mathbf{X})K_k^{cc}(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]} - m(\mathbf{x})\right| \\ &\leq C_1\left(1-\frac{1}{2d}\right)^k,\end{aligned}$$

where $C_1 = Ld$. Take $\alpha \in]0, 1/2]$. Let $\mathcal{C}_\alpha(\mathbf{x})$ be the event on which $\{|A_n(\mathbf{x})|, |B_n(\mathbf{x})| \leq \alpha\}$. On the event $\mathcal{C}_\alpha(\mathbf{x})$, we have

$$\begin{aligned}|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 &\leq 8|M_n(\mathbf{x}) - m(\mathbf{x})|^2 + 8|A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})|^2 \\ &\leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2.\end{aligned}$$

Thus,

$$\mathbb{E}[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}] \leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2. \tag{4.14}$$

Consequently, to find an upper bound on the rate of consistency of $\widetilde{m}^{cc}_{\infty,n}$, we just need to upper bound

$$\begin{aligned}\mathbb{E}\Big[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] &\leq \mathbb{E}\Big[|\max_{1\leq i\leq n} Y_i + m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \\ &\quad \text{(since } \widetilde{m}^{cc}_{\infty,n} \text{ is a local averaging estimate)} \\ &\leq \mathbb{E}\Big[|2\|m\|_\infty + \max_{1\leq i\leq n} \varepsilon_i|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \\ &\leq \left(\mathbb{E}\left[2\|m\|_\infty + \max_{1\leq i\leq n} \varepsilon_i\right]^4 \mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]\right)^{1/2} \\ &\quad \text{(by Cauchy-Schwarz inequality)} \\ &\leq \left(\left(16\|m\|^4_\infty + 8\mathbb{E}\Big[\max_{1\leq i\leq n} \varepsilon_i\Big]^4\right) \mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]\right)^{1/2}.\end{aligned}$$

Simple calculations on Gaussian tails show that one can find a constant $C > 0$ such that for all n,

$$\mathbb{E}\Big[\max_{1\leq i\leq n} \varepsilon_i\Big]^4 \leq C(\log n)^2.$$

Thus, there exists C_2 such that, for all $n > 1$,

$$\mathbb{E}\Big[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \leq C_2(\log n)(\mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right])^{1/2}. \tag{4.15}$$

The last probability $\mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]$ can be upper bounded by using Chebyshev's inequality. Indeed,

with respect to $A_n(\mathbf{x})$,

$$\begin{aligned}
\mathbb{P}\big[|A_n(\mathbf{x})| > \alpha\big] &\leq \frac{1}{n\alpha^2}\mathbb{E}\Big[\frac{Y K_k^{cc}(\mathbf{x},\mathbf{X})}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]} - \frac{\mathbb{E}\left[Y K_k^{cc}(\mathbf{x},\mathbf{X})\right]}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]}\Big]^2 \\
&\leq \frac{1}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right])^2}\mathbb{E}\Big[Y^2 K_k^{cc}(\mathbf{x},\mathbf{X})^2\Big] \\
&\leq \frac{2}{n\alpha^2}\frac{1}{(\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right])^2}\Big(\mathbb{E}\Big[m(\mathbf{X})^2 K_k^{cc}(\mathbf{x},\mathbf{X})^2\Big] \\
&\qquad + \mathbb{E}\Big[\varepsilon^2 K_k^{cc}(\mathbf{x},\mathbf{X})^2\Big]\Big) \\
&\leq \frac{2(\|m\|_\infty^2 + \sigma^2)}{n\alpha^2}\frac{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]}{(\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right])^2} \\
&\quad (\text{since} \sup_{\mathbf{x},\mathbf{z}\in[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) \leq 1) \\
&\leq \frac{2M_1^2}{\alpha^2}\frac{2^k}{n} \\
&\quad (\text{according to inequality (4.13)}),
\end{aligned}$$

where $M_1^2 = \|m\|_\infty^2 + \sigma^2$. Meanwhile with respect to $B_n(\mathbf{x})$, we obtain, still by Chebyshev's inequality,

$$\begin{aligned}
\mathbb{P}\big[|B_n(\mathbf{x})| > \alpha\big] &\leq \frac{1}{n\alpha^2}\mathbb{E}\Big[\frac{K_k^{cc}(\mathbf{x},\mathbf{X}_i)}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]}\Big]^2 \\
&\leq \frac{1}{n\alpha^2}\frac{1}{\mathbb{E}\left[K_k^{cc}(\mathbf{x},\mathbf{X})\right]} \\
&\quad (\text{since} \sup_{\mathbf{x},\mathbf{z}\in[0,1]^d} K_k^{cc}(\mathbf{x},\mathbf{z}) \leq 1) \\
&\leq \frac{2^k}{n\alpha^2}.
\end{aligned}$$

Thus, the probability of $\mathcal{C}_\alpha(\mathbf{x})$ is given by

$$\begin{aligned}
\mathbb{P}\big[\mathcal{C}_\alpha(\mathbf{x})\big] &\geq 1 - \mathbb{P}\big(|A_n(\mathbf{x})| \geq \alpha\big) - \mathbb{P}\big(|B_n(\mathbf{x})| \geq \alpha\big) \\
&\geq 1 - \frac{2^k}{n}\frac{2M_1^2}{\alpha^2} - \frac{2^k}{n\alpha^2} \\
&\geq 1 - \frac{2^k(2M_1^2+1)}{n\alpha^2}.
\end{aligned}$$

Consequently, according to inequality (4.15), we obtain

$$\mathbb{E}\Big[|\widetilde{m}_{\infty,n}^{cc}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\Big] \leq C_2(\log n)\Big(\frac{2^k(2M_1^2+1)}{n\alpha^2}\Big)^{1/2}.$$

Then using inequality (4.14),

$$\begin{aligned}
&\mathbb{E}\Big[\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\Big]^2 \\
&\leq \mathbb{E}\Big[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}\Big] + \mathbb{E}\Big[|\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \\
&\leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2 + C_2(\log n)\Big(\frac{2^k(2M_1^2+1)}{n\alpha^2}\Big)^{1/2}
\end{aligned}$$

Optimizing the right hand side in α, we get

$$\mathbb{E}\Big[\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\Big]^2 \leq 8C_1^2 \left(1 - \frac{1}{2d}\right)^{2k} + C_3\Big(\frac{(\log n)^2 2^k}{n}\Big)^{1/3},$$

for some constant $C_3 > 0$. The last expression is minimized for

$$k = C_4 + \frac{1}{\log 2 + \frac{3}{d}} \log\left(\frac{n}{(\log n)^2}\right),$$

where $C_4 = \left(\frac{1}{d} + \frac{\log 2}{3}\right)^{-1} \log\left(\frac{C_3 d \log 2}{24 C_1^2}\right)$. Consequently, there exists a constant C_5 such that, for all $n > 1$,

$$\mathbb{E}\left[\widetilde{m}^{cc}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq C_5 n^{-\frac{1}{d\log 2 + 3}} (\log n)^2.$$

□

Proof of Lemma 4.1. Let $x, z \in [0,1]$ such that $x < z$. The first statement comes from the fact that splits are drawn uniformly over $[0,1]$. To address the second one, denote by Z_1 (resp. Z_2) the position of the first (resp. second) split used to build the cell containing $[x,z]$. Observe that, given $Z_1 = z_1$, Z_2 is uniformly distributed over $[z_1, 1]$ (resp. $[0, z_1]$) if $z_1 \leq x$ (resp. $z_1 \geq z$). Thus, we have

$$\begin{aligned}
K_2^{uf}(x,z) = &\int_{z_1=0}^{x} \left(\int_{z_2=z_1}^{x} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 + \int_{z_2=z}^{1} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2\right) \\
&+ \int_{z_1=z}^{1} \left(\int_{z_2=0}^{x} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 + \int_{z_2=z}^{z_1} \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2\right).
\end{aligned}$$

The first term takes the form

$$\begin{aligned}
\int_0^x \frac{1}{z_1} \left(\int_{z_1}^{x} \mathrm{d}z_2\right) \mathrm{d}z_1 &= \int_0^x \frac{x - z_1}{1 - z_1} \mathrm{d}z_1 \\
&= x - (1-x)\log(1-x).
\end{aligned}$$

Similarly, one has

$$\begin{aligned}
\int_0^x \int_z^1 \frac{1}{1-z_1} \mathrm{d}z_1 \mathrm{d}z_2 &= (1-z)\log(1-x), \\
\int_z^1 \int_z^{z_1} \frac{1}{z_1} \mathrm{d}z_1 \mathrm{d}z_2 &= (1-z) + z\log z, \\
\int_z^1 \int_0^x \frac{1}{z_1} \mathrm{d}z_1 \mathrm{d}z_2 &= -x \log z.
\end{aligned}$$

Consequently,

$$\begin{aligned} K_2^{uf}(x,z) &= x - (1-x)\log(1-x) + (1-z)\log(1-x) \\ &\quad - x\log z + (1-z) + z\log z \\ &= 1 - (z-x) + (z-x)\log\left(\frac{z}{1-x}\right). \end{aligned}$$

□

Proof of Proposition 4.6. The result is proved in Technical Proposition 2 in Scornet [2014].

□

To prove Theorem 4.2, we need to control the bias of uniform KeRF estimates, which is done in Theorem 4.4.

Theorem 4.4. *Let f be a L-Lipschitz function. Then, for all k,*

$$\sup_{\mathbf{x}\in[0,1]^d} \left| \frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|) f(\mathbf{z}) \mathrm{d}z_1 \dots \mathrm{d}z_d}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|) \mathrm{d}z_1 \dots \mathrm{d}z_d} - f(\mathbf{x}) \right| \leq \frac{Ld2^{2d+1}}{3} \left(1 - \frac{1}{3d}\right)^k.$$

Proof of Theorem 4.4. Let $\mathbf{x} \in [0,1]^d$ and $k \in \mathbb{N}$. Let f be a L-Lipschitz function. In the rest of the proof, for clarity reasons, we use the notation $\mathrm{d}\mathbf{z}$ instead of $\mathrm{d}z_1 \dots \mathrm{d}z_d$. Thus,

$$\left| \frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|) f(\mathbf{z}) \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|) \mathrm{d}\mathbf{z}} - f(\mathbf{x}) \right| \leq \frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|) |f(\mathbf{z}) - f(\mathbf{x})| \mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|) \mathrm{d}\mathbf{z}}.$$

Note that,

$$
\begin{aligned}
&\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)|f(\mathbf{z})-f(\mathbf{x})|\mathrm{d}\mathbf{z} \\
&\leq L\sum_{\ell=1}^{d}\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)|z_\ell - x_\ell|\mathrm{d}\mathbf{z} \\
&\leq L\sum_{\ell=1}^{d}\int_{[0,1]^d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{d}\right)^k \prod_{m=1}^{d} K_{k_m}^{uf}(0,|z_m - x_m|)|z_\ell - x_\ell|\mathrm{d}\mathbf{z} \\
&\leq L\sum_{\ell=1}^{d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{1}{d}\right)^k \prod_{m\neq\ell}\int_0^1 K_{k_m}^{uf}(0,|z_m - x_m|)\mathrm{d}z_m \\
&\qquad\qquad \times \int_0^1 K_{k_\ell}^{uf}(0,|z_\ell - x_\ell|)|z_\ell - x_\ell|\mathrm{d}z_\ell \\
&\leq L\sum_{\ell=1}^{d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{2}{3}\right)^{k_l+1}\left(\frac{1}{d}\right)^k \prod_{m=1}^{d}\int_0^1 K_{k_m}^{uf}(0,|z_m - x_m|)\mathrm{d}z_m \\
&\qquad \text{(according to the second statement of Lemma 4.2, see below)} \\
&\leq \frac{L}{2^{k-d}}\sum_{\ell=1}^{d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{2}{3}\right)^{k_\ell+1}\left(\frac{1}{d}\right)^k, \qquad (4.16)
\end{aligned}
$$

according to the first statement of Lemma 4.2. Still by Lemma 4.2 and using inequality (4.16), we have,

$$
\begin{aligned}
&\frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)|f(\mathbf{z})-f(\mathbf{x})|\mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)\mathrm{d}\mathbf{z}} \\
&\qquad \leq \frac{L2^{2d+1}}{3}\sum_{\ell=1}^{d} \sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{2}{3}\right)^{k_\ell}\left(\frac{1}{d}\right)^k.
\end{aligned}
$$

Taking the first term of the sum, we obtain

$$
\begin{aligned}
\sum_{\substack{k_1,\ldots,k_d \\ \sum_{j=1}^d k_j = k}} \frac{k!}{k_1!\ldots k_d!}\left(\frac{2}{3}\right)^{k_1}\left(\frac{1}{d}\right)^k &= \sum_{k_1=0}^{k}\left(\frac{2}{3d}\right)^{k_1}\left(1-\frac{1}{d}\right)^{k-k_1}\frac{k!}{k_1!(k-k_1)!} \\
&\leq \left(1-\frac{1}{3d}\right)^k.
\end{aligned}
$$

Finally,

$$\frac{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)|f(\mathbf{z})-f(\mathbf{x})|\mathrm{d}\mathbf{z}}{\int_{[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z}-\mathbf{x}|)\mathrm{d}\mathbf{z}} \leq \frac{L2^{2d+1}}{3}\left(1-\frac{1}{3d}\right)^k.$$

□

Proof of Theorem 4.2. Let $\mathbf{x} \in [0,1]^d$, $\|m\|_\infty = \sup_{\mathbf{x}\in[0,1]^d} |m(\mathbf{x})|$ and recall that

$$m_{\infty,n}^{uf}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\sum_{i=1}^n K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}.$$

Thus, letting

$$\begin{aligned}
A_n(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} - \frac{\mathbb{E}[Y K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} \right), \\
B_n(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^n \left(\frac{K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} - 1 \right), \\
\text{and } M_n(\mathbf{x}) &= \frac{\mathbb{E}[Y K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]},
\end{aligned}$$

the estimate $m_{\infty,n}^{uf}(\mathbf{x})$ can be rewritten as

$$m_{\infty,n}^{uf}(\mathbf{x}) = \frac{M_n(\mathbf{x}) + A_n(\mathbf{x})}{1 + B_n(\mathbf{x})},$$

which leads to

$$m_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x}) = \frac{M_n(\mathbf{x}) - m(\mathbf{x}) + A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})}{1 + B_n(\mathbf{x})}.$$

Note that, according to Theorem 4.4, we have

$$\begin{aligned}
|M_n(\mathbf{x}) - m(\mathbf{x})| &= \left| \frac{\mathbb{E}[m(\mathbf{X})K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} + \frac{\mathbb{E}[\varepsilon K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} - m(\mathbf{x}) \right| \\
&\leq \left| \frac{\mathbb{E}[m(\mathbf{X})K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X}-\mathbf{x}|)]} - m(\mathbf{x}) \right| \\
&\leq C_1 \left(1 - \frac{1}{3d}\right)^k,
\end{aligned}$$

where $C_1 = L2^{2d+1}/3$. Take $\alpha \in]0, 1/2]$. Let $\mathcal{C}_\alpha(\mathbf{x})$ be the event on which $\{|A_n(\mathbf{x})|, |B_n(\mathbf{x})| \leq \alpha\}$. On the event $\mathcal{C}_\alpha(\mathbf{x})$, we have

$$\begin{aligned}
|m_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 &\leq 8|M_n(\mathbf{x}) - m(\mathbf{x})|^2 + 8|A_n(\mathbf{x}) - B_n(\mathbf{x})m(\mathbf{x})|^2 \\
&\leq 8C_1^2 \left(1 - \frac{1}{3d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2.
\end{aligned}$$

Thus,

$$\mathbb{E}[|m^{uf}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}] \leq 8C_1^2 \left(1 - \frac{1}{3d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2. \tag{4.17}$$

Consequently, to find an upper bound on the rate of consistency of $m^{uf}_{\infty,n}$, we just need to upper bound

$$\begin{aligned}
\mathbb{E}\Big[|\widetilde{m}^{uf}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] &\leq \mathbb{E}\Big[|\max_{1\leq i\leq n} Y_i + m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \\
&\quad \text{(since } \widetilde{m}^{uf}_{\infty,n} \text{ is a local averaging estimate)} \\
&\leq \mathbb{E}\Big[|2\|m\|_\infty + \max_{1\leq i\leq n} \varepsilon_i|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \\
&\leq \left(\mathbb{E}\left[2\|m\|_\infty + \max_{1\leq i\leq n} \varepsilon_i\right]^4 \mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]\right)^{1/2} \\
&\quad \text{(by Cauchy-Schwarz inequality)} \\
&\leq \left(\left(16\|m\|^4_\infty + 8\mathbb{E}\Big[\max_{1\leq i\leq n} \varepsilon_i\Big]^4\right) \mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]\right)^{1/2}.
\end{aligned}$$

Simple calculations on Gaussian tails show that one can find a constant $C > 0$ such that for all n,

$$\mathbb{E}\Big[\max_{1\leq i\leq n} \varepsilon_i\Big]^4 \leq C(\log n)^2.$$

Thus, there exists C_2 such that, for all $n > 1$,

$$\mathbb{E}\Big[|\widetilde{m}^{uf}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}^c_\alpha(\mathbf{x})}\Big] \leq C_2(\log n)(\mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right])^{1/2}. \tag{4.18}$$

The last probability $\mathbb{P}\left[\mathcal{C}^c_\alpha(\mathbf{x})\right]$ can be upper bounded by using Chebyshev's inequality. Indeed, with respect to $A_n(\mathbf{x})$,

$$\begin{aligned}
\mathbb{P}\left[|A_n(\mathbf{x})| > \alpha\right] &\leq \frac{1}{n\alpha^2}\mathbb{E}\left[\frac{Y K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)}{\mathbb{E}[K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]} - \frac{\mathbb{E}[Y K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}{\mathbb{E}[K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}\right]^2 \\
&\leq \frac{1}{n\alpha^2} \frac{1}{(\mathbb{E}[K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)])^2} \mathbb{E}\Big[Y^2 K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)^2\Big] \\
&\leq \frac{2}{n\alpha^2} \frac{1}{(\mathbb{E}[K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)])^2} \Big(\mathbb{E}[m(\mathbf{X})^2 K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)^2] \\
&\qquad + \mathbb{E}[\varepsilon^2 K^{uf}_k(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)^2]\Big),
\end{aligned}$$

which leads to

$$\begin{aligned}\mathbb{P}\big[|A_n(\mathbf{x})| > \alpha\big] &\leq \frac{2(\|m\|_\infty^2 + \sigma^2)}{n\alpha^2} \frac{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}{(\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)])^2} \\ &\quad \text{(since } \sup_{\mathbf{x},\mathbf{z}\in[0,1]^d} K_k^{uf}(\mathbf{0}, |\mathbf{z} - \mathbf{x}|) \leq 1) \\ &\leq \frac{M_1^2}{\alpha^2}\frac{2^k}{n} \\ &\quad \text{(according to the first statement of Lemma 4.2),}\end{aligned}$$

where $M_1^2 = 2^{d+1}(\|m\|_\infty^2 + \sigma^2)$. Meanwhile with respect to $B_n(\mathbf{x})$, we have, still by Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}\big[|B_n(\mathbf{x})| > \alpha\big] &\leq \frac{1}{n\alpha^2}\mathbb{E}\left[\frac{K_k^{uf}(\mathbf{0}, |\mathbf{X}_i - \mathbf{x}|)}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]}\right]^2 \\ &\leq \frac{1}{n\alpha^2}\frac{1}{\mathbb{E}[K_k^{uf}(\mathbf{0}, |\mathbf{X} - \mathbf{x}|)]} \\ &\leq \frac{2^{k+d}}{n\alpha^2}.\end{aligned}$$

Thus, the probability of $\mathcal{C}_\alpha(\mathbf{x})$ is given by

$$\begin{aligned}\mathbb{P}\big[\mathcal{C}_\alpha(\mathbf{x})\big] &\geq 1 - \mathbb{P}\big(|A_n(\mathbf{x})| \geq \alpha\big) - \mathbb{P}\big(|B_n(\mathbf{x})| \geq \alpha\big) \\ &\geq 1 - \frac{2^k}{n}\frac{M_1^2}{\alpha^2} - \frac{2^{k+d}}{n\alpha^2} \\ &\geq 1 - \frac{2^k(M_1^2 + 2^d)}{n\alpha^2}.\end{aligned}$$

Consequently, according to inequality (4.18), we obtain

$$\mathbb{E}\Big[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\Big] \leq C_2(\log n)\left(\frac{2^k(M_1^2 + 2^d)}{n\alpha^2}\right)^{1/2}.$$

Then using inequality (4.17),

$$\begin{aligned}&\mathbb{E}\Big[\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})\Big]^2 \\ &\leq \mathbb{E}\Big[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha(\mathbf{x})}\Big] + \mathbb{E}\Big[|\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbb{1}_{\mathcal{C}_\alpha^c(\mathbf{x})}\Big] \\ &\leq 8C_1^2\left(1 - \frac{1}{3d}\right)^{2k} + 8\alpha^2(1 + \|m\|_\infty)^2 + C_2(\log n)\left(\frac{2^k(M_1^2 + 2^d)}{n\alpha^2}\right)^{1/2}.\end{aligned}$$

Optimizing the right hand side in α, we get

$$\mathbb{E}\Big[\widetilde{m}_{\infty,n}^{uf}(\mathbf{x}) - m(\mathbf{x})\Big]^2 \leq 8C_1^2\left(1 - \frac{1}{3d}\right)^{2k} + C_3\left(\frac{(\log n)^2 2^k}{n}\right)^{1/3},$$

for some constant $C_3 > 0$. The last expression is minimized for

$$k = C_4 + \frac{1}{\log 2 + \frac{2}{d}} \log\left(\frac{n}{(\log n)^2}\right),$$

where $C_4 = -3\Big(\log 2 + \frac{2}{d}\Big)^{-1} \log\left(\frac{dC_3 \log 2}{16 C_1^2}\right)$. Thus, there exists a constant $C_5 > 0$ such that, for all $n > 1$,

$$\mathbb{E}\left[\widetilde{m}^{uf}_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\right]^2 \leq C n^{-2/(6+3d\log 2)} (\log n)^2.$$

□

Lemma 4.2. *For all $k \in \mathbb{N}$ and $x \in [0,1]$,*

(i)

$$\left(\frac{1}{2}\right)^{k_l+1} \leq \int_0^1 K^{uf}_{k_l}(0, |z_l - x_l|)\mathrm{d}z \leq \left(\frac{1}{2}\right)^{k_l-1}.$$

(ii)

$$\int_{[0,1]} K^{uf}_{k_l}(0, |z_l - x_l|)|x_l - z_l| dz_l \leq \left(\frac{2}{3}\right)^{k_l+1} \int_{[0,1]} K^{uf}_{k_l}(0, |z_l - x_l|) dz_l.$$

Proof of Lemma 4.2. Let $k_l \in \mathbb{N}$ and $x_l \in [0,1]$. We start by proving (i). According to Proposition 4.6, the connection function of uniform random forests of level k_l takes the form

$$\begin{aligned}
\int_{[0,1]} K^{uf}_{k_l}(0, |z_l - x_l|) dz_l &= \int_{-\log x_l}^{\infty} e^{-2u} \sum_{j=k_l}^{\infty} \frac{u^j}{j!}\mathrm{d}u + \int_{-\log(1-x_l)}^{\infty} e^{-2u} \sum_{j=k_l}^{\infty} \frac{u^j}{j!}\mathrm{d}u \\
&= \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} \int_{-2\log x_l}^{\infty} e^{-u} \frac{u^j}{j!}\mathrm{d}u \\
&\quad + \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} \int_{-2\log(1-x_l)}^{\infty} e^{-u} \frac{u^j}{j!}\mathrm{d}u \\
&= \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} x_l^2 \sum_{i=0}^{j} \frac{(-2\log x_l)^i}{i!} \\
&\quad + \sum_{j=k_l}^{\infty} \left(\frac{1}{2}\right)^{j+1} (1-x_l)^2 \sum_{i=0}^{j} \frac{(-2\log(1-x_l))^i}{i!}.
\end{aligned}$$

Therefore,

$$\int_{[0,1]} K^{uf}_{k_l}(0, |z_l - x_l|)\mathrm{d}z_l \leq \left(\frac{1}{2}\right)^{k_l-1},$$

and

$$\int_{[0,1]} K_{k_l}^{uf}(0,|z_l-x_l|)\mathrm{d}z_l \geq \left(x_l^2+(1-x_l)^2\right)\left(\frac{1}{2}\right)^{k_l} \geq \left(\frac{1}{2}\right)^{k_l+1}.$$

Regarding the second statement of Lemma 4.2, we have

$$\begin{aligned}
&\int_{[0,1]} K_{k_l}^{uf}(0,|z_l-x_l|)|x_l-z_l|\mathrm{d}z_l \\
&= \int_{[0,1]} |x_l-z_l|^2 \sum_{j=k_l}^{\infty} \frac{(-\log|x_l-z_l|)^j}{j!}\mathrm{d}z_l \\
&= \int_{z_l\leq x_l} (x_l-z_l)^2 \sum_{j=k_l}^{\infty} \frac{(-\log|x_l-z_l|)^j}{j!}\mathrm{d}z_l \\
&\quad + \int_{z_l>x_l} (z_l-x_l)^2 \sum_{j=k_l}^{\infty} \frac{(-\log|x_l-z_l|)^j}{j!}\mathrm{d}z_l \\
&= \int_{[0,x_l]} v^2 \sum_{j=k_l}^{\infty} \frac{(-\log v)^j}{j!}dv + \int_{[0,1-x_l]} u^2 \sum_{j=k_l}^{\infty} \frac{(-\log u)^j}{j!}du \\
&= \int_{-\log(x_l)}^{\infty} e^{-3w} \sum_{j=k_l}^{\infty} \frac{w^j}{j!}\mathrm{d}w + \int_{-\log(1-x_l)}^{\infty} e^{-3w} \sum_{j=k_l}^{\infty} \frac{w^j}{j!}\mathrm{d}w \\
&= \frac{2}{3}\int_{-3\log(x_l)/2}^{\infty} e^{-2w} \sum_{j=k_l}^{\infty} \frac{(2v/3)^j}{j!}dv + \frac{2}{3}\int_{-3\log(1-x_l)/2}^{\infty} e^{-2v} \sum_{j=k_l}^{\infty} \frac{(2v/3)^j}{j!}dv \\
&\leq \left(\frac{2}{3}\right)^{k_l+1}\left(\int_{-\log(x_l)}^{\infty} e^{-2w} \sum_{j=k_l}^{\infty} \frac{v^j}{j!}dv + \int_{-\log(1-x_l)}^{\infty} e^{-2v} \sum_{j=k_l}^{\infty} \frac{v^j}{j!}\mathrm{d}v\right) \\
&\leq \left(\frac{2}{3}\right)^{k_l+1} \int_{[0,1]} K_{k_l}^{uf}(0,|z_l-x_l|)\mathrm{d}z_l.
\end{aligned}$$

□

Chapter 5

Consistency of random forests

Abstract Random forests are a learning algorithm proposed by Breiman [2001] that combines several randomized decision trees and aggregates their predictions by averaging. Despite its wide usage and outstanding practical performance, little is known about the mathematical properties of the procedure. This disparity between theory and practice originates in the difficulty to simultaneously analyze both the randomization process and the highly data-dependent tree structure. In the present paper, we take a step forward in forest exploration by proving a consistency result for Breiman's [2001] original algorithm in the context of additive regression models. Our analysis also sheds an interesting light on how random forests can nicely adapt to sparsity.

We greatly thank two referees for valuable comments and insightful suggestions. This work was supported by the European Research Council [SMAC-ERC-280032].

Contents

5.1 Introduction

Random forests are an ensemble learning method for classification and regression that constructs a number of randomized decision trees during the training phase and predicts by averaging the results. Since its publication in the seminal paper of Breiman [2001], the procedure has become a major data analysis tool, that performs well in practice in comparison with many standard methods. What has greatly contributed to the popularity of forests is the fact that they can be applied to a wide range of prediction problems and have few parameters to tune. Aside from being simple to use, the method is generally recognized for its accuracy and its ability to deal with small sample sizes, high-dimensional feature spaces, and complex data structures. The random forest methodology has been successfully involved in many practical problems, including air quality prediction (winning code of the EMC data science global hackathon in 2012, see `http://www.kaggle.com/c/dsg-hackathon`), chemoinformatics [Svetnik et al., 2003], ecology [Prasad et al., 2006, Cutler et al., 2007], 3D object recognition [Shotton et al., 2011], and bioinformatics [Díaz-Uriarte and de Andrés, 2006], just to name a few. In addition, many variations on the original algorithm have been proposed to improve the calculation time while maintaining good prediction accuracy [see, e.g., Geurts et al., 2006, Amaratunga et al., 2008]. Breiman's forests have also been extended to quantile estimation [Meinshausen, 2006], survival analysis [Ishwaran et al., 2008], and ranking prediction [Clémençon et al., 2013].

On the theoretical side, the story is less conclusive and, regardless of their extensive use in practical settings, little is known about the mathematical properties of random forests. To date, most studies have concentrated on isolated parts or simplified versions of the procedure. The most celebrated theoretical result is that of Breiman [2001], which offers an upper bound on the generalization error of forests in terms of correlation and strength of the individual trees. This was followed by a technical note [Breiman, 2004], that focuses on a stylized version of the original algorithm. A critical step was subsequently taken by Lin and Jeon [2006], who established lower bounds for non-adaptive forests (i.e., independent of the training set). They also highlighted an interesting connection between random forests and a particular class of nearest neighbor predictors that was further worked out by Biau and Devroye [2010]. In recent years, various theoretical studies [e.g., Biau et al., 2008, Ishwaran and Kogalur, 2010, Biau, 2012, Genuer, 2012, Zhu et al., 2012] have been performed, analyzing consistency of simplified models, and moving ever closer to practice. Recent attempts towards narrowing the gap between theory and practice are by Denil et al. [2013], who proves the first consistency result for online random forests, and by Wager [2014] and Mentch and Hooker [2014a] who study the asymptotic sampling distribution of forests.

The difficulty to properly analyze random forests can be explained by the black-box nature of the procedure, which is actually a subtle combination of different components. Among the forest essential ingredients, both bagging [Breiman, 1996] and the Classification And Regression Trees (CART)-split criterion [Breiman et al., 1984] play a critical role. Bagging (a contraction of bootstrap-aggregating) is a general aggregation scheme which proceeds by generating subsamples from the original data set, constructing a predictor from each resample and deciding by averaging. It is one of the most effective computationally intensive procedures to improve on unstable estimates, especially for large, high-dimensional data sets where finding a good model in one step is impossible because of the complexity and scale of the problem [Bühlmann and Yu,

2002, Kleiner et al., 2012, Wager et al., 2013]. As for the CART-split selection, it is originated from the most influential CART algorithm of Breiman et al. [1984], and is used in the construction of the individual trees to choose the best cuts perpendicular to the axes. At each node of each tree, the best cut is selected by optimizing the CART-split criterion, based on the notion of Gini impurity (classification) and prediction squared error (regression).

Yet, while bagging and the CART-splitting scheme play a key role in the random forest mechanism, both are difficult to analyze, thereby explaining why theoretical studies have considered so far simplified versions of the original procedure. This is often done by simply ignoring the bagging step and by replacing the CART-split selection by a more elementary cut protocol. Besides, in Breiman's forests, each leaf (that is, a terminal node) of the individual trees contains a fixed pre-specified number of observations (this parameter, called `nodesize` in the R package `randomForests`, is usually chosen between 1 and 5). There is also an extra parameter in the algorithm which allows to control the total number of leaves (this parameter is called `maxnode` in the R package and has, by default, no effect on the procedure). The combination of these various components makes the algorithm difficult to analyze with rigorous mathematics. As a matter of fact, most authors focus on simplified, data-independent procedures, thus creating a gap between theory and practice.

Motivated by the above discussion, we study in the present paper some asymptotic properties of Breiman's [2001] algorithm in the context of additive regression models. We prove the $\mathbb{L}^2$ consistency of random forests, which gives a first basic theoretical guarantee of efficiency for this algorithm. Up to our knowledge, this is the first consistency result for Breiman's [2001] original procedure. Our approach rests upon a detailed analysis of the behavior of the cells generated by CART-split selection as the sample size grows. It turns out that a good control of the regression function variation inside each cell, together with a proper choice of the total number of leaves (Theorem 5.1) or a proper choice of the subsampling rate (Theorem 5.2) are sufficient to ensure the forest consistency in a $\mathbb{L}^2$ sense. Also, our analysis shows that random forests can adapt to a sparse framework, when the ambient dimension p is large (independent of n), but only a smaller number of coordinates carry out information.

The paper is organized as follows. In Section 2, we introduce some notations and describe the random forest method. The main asymptotic results are presented in Section 3 and further discussed in Section 4. Section 5 is devoted to the main proofs, and technical results are gathered in the supplemental article [Scornet et al., 2015a].

5.2 Random forests

The general framework is $\mathbb{L}^2$ regression estimation, in which an input random vector $\mathbf{X} \in [0,1]^p$ is observed, and the goal is to predict the square integrable random response $Y \in \mathbb{R}$ by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. To this aim, we assume given a training sample $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of $[0,1]^p \times \mathbb{R}$-valued independent random variables distributed as the independent prototype pair $(\mathbf{X}, Y)$. The objective is to use the data set $\mathcal{D}_n$ to construct an estimate $m_n : [0,1]^p \to \mathbb{R}$ of the function m. In this respect, we say that a regression function estimate m_n is $\mathbb{L}^2$ consistent if $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \to 0$ as $n \to \infty$ (where the expectation is over $\mathbf{X}$ and $\mathcal{D}_n$).

A random forest is a predictor consisting of a collection of M randomized regression trees. For the j-th tree in the family, the predicted value at the query point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ and independent of $\mathcal{D}_n$. In practice, this variable is used to resample the training set prior to the growing of individual trees and to select the successive candidate directions for splitting. The trees are combined to form the (finite) forest estimate

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n). \tag{5.1}$$

Since in practice we can choose M as large as possible, we study in this paper the property of the infinite forest estimate obtained as the limit of (5.1) when the number of trees M grows to infinity:

$$m_{\infty,n}(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}; \Theta, \mathcal{D}_n) \right],$$

where $\mathbb{E}_\Theta$ denotes expectation with respect to the random parameter Θ, conditional on $\mathcal{D}_n$. This operation is justified by the law of large numbers, which asserts that, almost surely, conditional on $\mathcal{D}_n$,

$$\lim_{M \to \infty} m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = m_{\infty,n}(\mathbf{x}; \mathcal{D}_n)$$

[see, e.g., Breiman, 2001, Scornet, 2014, for details]. In the sequel, to lighten notation, we will simply write $m_{\infty,n}(\mathbf{x})$ instead of $m_{\infty,n}(\mathbf{x}; \mathcal{D}_n)$.

In Breiman's [2001] original forests, each node of a single tree is associated with a hyper-rectangular cell. At each step of the tree construction, the collection of cells forms a partition of $[0,1]^p$. The root of the tree is $[0,1]^p$ itself, and each tree is grown as explained in Algorithm 4.

This algorithm has three parameters:

1. $m_{\text{try}} \in \{1, \ldots, p\}$, which is the number of pre-selected directions for splitting;
2. $a_n \in \{1, \ldots, n\}$, which is the number of sampled data points in each tree;
3. $t_n \in \{1, \ldots, a_n\}$, which is the number of leaves in each tree.

By default, in the original procedure, the parameter m_{try} is set to $p/3$, a_n is set to n (resampling is done with replacement), and $t_n = a_n$. However, in our approach, resampling is done without replacement and the parameters a_n and t_n can be different from their default values.

In a word, the algorithm works by growing M different trees as follows. For each tree, a_n data points are drawn at random without replacement from the original data set; then, at each cell of every tree, a split is chosen by maximizing the CART-criterion (see below); finally, the construction of every tree is stopped when the total number of cells in the tree reaches the value t_n (therefore, each cell contains exactly one point in the case $t_n = a_n$).

Algorithm 4: Breiman's random forest predicted value at $\mathbf{x}$.

Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $m_{\text{try}} \in \{1, \dots, p\}$, $a_n \in \{1, \dots, n\}$, $t_n \in \{1, \dots, a_n\}$, and $\mathbf{x} \in [0,1]^p$.

Output: Prediction of the random forest at $\mathbf{x}$.

for $j = 1, \dots, M$ **do**
  Select a_n points, without replacement, uniformly in $\mathcal{D}_n$.
  Set $\mathcal{P}_0 = \{[0,1]^p\}$ the partition associated with the root of the tree.
  For all $1 \leq \ell \leq a_n$, set $\mathcal{P}_\ell = \emptyset$.
  Set $n_{nodes} = 1$ and $level = 0$.
  while $n_{nodes} < t_n$ **do**
    if $\mathcal{P}_{level} = \emptyset$ **then**
      $level = level + 1$
    else
      Let A be the first element in $\mathcal{P}_{level}$.
      if A *contains exactly one point* **then**
        $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$
        $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A\}$
      else
        Select uniformly, without replacement, a subset $\mathcal{M}_{\text{try}} \subset \{1, \dots, p\}$ of cardinality m_{try}.
        Select the best split in A by optimizing the CART-split criterion along the coordinates in $\mathcal{M}_{\text{try}}$ *(see details below)*.
        Cut the cell A according to the best split. Call A_L and A_R the two resulting cell.
        $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$
        $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A_L\} \cup \{A_R\}$
        $n_{nodes} = n_{nodes} + 1$
      end
    end
  end
  Compute the predicted value $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ at $\mathbf{x}$ equal to the average of the Y_i's falling in the cell of $\mathbf{x}$ in partition $\mathcal{P}_{level} \cup \mathcal{P}_{level+1}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n)$ at the query point $\mathbf{x}$ according to (5.1).

We note that the resampling step in Algorithm 4 (line 2) is done by choosing a_n out of n points (with $a_n \leq n$) without replacement. This is slightly different from the original algorithm, where resampling is done by bootstrapping, that is by choosing n out of n data points with replacement.

Selecting the points "without replacement" instead of "with replacement" is harmless—in fact, it is just a means to avoid mathematical difficulties induced by the bootstrap [see, e.g., Efron, 1982, Politis et al., 1999].

On the other hand, letting the parameters a_n and t_n depend upon n offers several degrees of freedom which opens the route for establishing consistency of the method. To be precise, we will study in Section 3 the random forest algorithm in two different regimes. The first regime is when $t_n < a_n$, which means that trees are not fully developed. In that case, a proper tuning of t_n ensures the forest's consistency (Theorem 5.1). The second regime occurs when $t_n = a_n$, i.e. when trees are fully grown. In that case, consistency results from an appropriate choice of the subsample rate a_n/n (Theorem 5.2).

So far, we have not made explicit the CART-split criterion used in Algorithm 4. To properly define it, we let A be a generic cell and $N_n(A)$ be the number of data points falling in A. A cut in A is a pair (j, z), where j is a dimension in $\{1, \ldots, p\}$ and z is the position of the cut along the j-th coordinate, within the limits of A. We let $\mathcal{C}_A$ be the set of all such possible cuts in A. Then, with the notation $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \ldots, \mathbf{X}_i^{(p)})$, for any $(j, z) \in \mathcal{C}_A$, the CART-split criterion [Breiman et al., 1984] takes the form

$$\begin{aligned} L_n(j,z) =& \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\ & - \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z})^2 \mathbb{1}_{\mathbf{X}_i \in A}, \end{aligned} \tag{5.2}$$

where $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$, and $\bar{Y}_A$ (resp., $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) is the average of the Y_i's belonging to A (resp., A_L, A_R), with the convention $0/0 = 0$. At each cell A, the best cut $(j_n^\star, z_n^\star)$ is finally selected by maximizing $L_n(j, z)$ over $\mathcal{M}_{\text{try}}$ and $\mathcal{C}_A$, that is

$$(j_n^\star, z_n^\star) \in \operatorname*{arg\,max}_{\substack{j \in \mathcal{M}_{\text{try}} \\ (j,z) \in \mathcal{C}_A}} L_n(j, z).$$

To remove ties in the argmax, the best cut is always performed along the best cut direction $j_n^\star$, at the middle of two consecutive data points.

5.3 Main results

We consider an additive regression model satisfying the following properties:

(H5.1) *The response Y follows*

$$Y = \sum_{j=1}^{p} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

where $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)})$ *is uniformly distributed over* $[0,1]^p$*,* ε *is an independent centered Gaussian noise with finite variance* $\sigma^2 > 0$*, and each component* m_j *is continuous.*

Additive regression models, which extend linear models, were popularized by Stone [1985] and Hastie and Tibshirani [1986]. These models, which decompose the regression function as a sum of univariate functions, are flexible and easy to interpret. They are acknowledged for providing a good trade-off between model complexity and calculation time, and were accordingly extensively studied for the last thirty years. Additive models also play an important role in the context of high-dimensional data analysis and sparse modelling, where they are successfully involved in procedures such as the Lasso and various aggregation schemes [for an overview, see, e.g., Hastie et al., 2009]. Although random forests fall in the family of non parametric procedures, it turns out that the analysis of their properties is facilitated within the framework of additive models.

Our first result assumes that the total number of leaves t_n in each tree tends to infinity slower than the number of selected data points a_n.

Theorem 5.1. *Assume that* **(H5.1)** *is satisfied. Then, provided* $a_n \to \infty$ *and* $t_n(\log a_n)^9/a_n \to 0$*, random forests are consistent, i.e.,*

$$\lim_{n\to\infty} \mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

Noteworthy, Theorem 5.1 still holds with $a_n = n$. In that case, the subsampling step plays no role in the consistency of the method. Indeed, controlling the depth of the trees via the parameter t_n is sufficient to bound the forest error. We note in passing that an easy adaptation of Theorem 5.1 shows that the CART algorithm is consistent under the same assumptions.

The term $(\log a_n)^9$ originates from the Gaussian noise and allows to control the noise tail. In the easier situation where the Gaussian noise is replaced by a bounded random variable, it is easy to see that the term $(\log a_n)^9$ turns into $\log a_n$, a term which accounts for the complexity of the tree partition.

Let us now examine the forest behavior in the second regime, where $t_n = a_n$ (i.e., trees are fully grown) and, as before, subsampling is done at the rate a_n/n. The analysis of this regime turns out to be more complicated, and rests upon assumption **(H5.2)** below. We denote by $Z_i = \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}$ the indicator that $\mathbf{X}_i$ falls in the same cell as $\mathbf{X}$ in the random tree designed with $\mathcal{D}_n$ and the random parameter Θ. Similarly, we let $Z'_j = \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}$, where Θ' is an independent copy of Θ. Accordingly, we define

$$\psi_{i,j}(Y_i, Y_j) = \mathbb{E}\Big[Z_i Z'_j | \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i, Y_j\Big]$$

$$\text{and} \quad \psi_{i,j} = \mathbb{E}\Big[Z_i Z'_j | \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n\Big].$$

Finally, for any random variables W_1, W_2, Z, we denote by $\mathrm{Corr}(W_1, W_2|Z)$ the conditional correlation coefficient (whenever it exists).

(H5.2) *Let* $Z_{i,j} = (Z_i, Z'_j)$*. Then, one of the following two conditions holds:*

(a) *One has*

$$\lim_{n\to\infty} (\log a_n)^{2p-2} (\log n)^2 \mathbb{E}\left[\max_{\substack{i,j \\ i\neq j}} |\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}|\right]^2 = 0.$$

(b) *There exist a constant $C > 0$ and a sequence $(\gamma_n)_n \to 0$ such that, almost surely,*

$$\max_{\ell_1,\ell_2=0,1} \frac{|\operatorname{Corr}(Y_i - m(\mathbf{X}_i), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)} | \mathbf{X}_i, \mathbf{X}_j, Y_j)|}{\mathbb{P}^{1/2}[Z_{i,j} = (\ell_1,\ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j]} \leq \gamma_n,$$

and

$$\max_{\ell_1=0,1} \frac{|\operatorname{Corr}\left((Y_i - m(\mathbf{X}_i))^2, \mathbb{1}_{Z_i=\ell_1} | \mathbf{X}_i\right)|}{\mathbb{P}^{1/2}[Z_i = \ell_1|\mathbf{X}_i]} \leq C.$$

Despite their technical aspect, statements **(H5.2a)** and **(H5.2b)** have simple interpretations. To understand the meaning of **(H5.2a)**, let us replace the Gaussian noise by a bounded random variable. A close inspection of Lemma 5.4 shows that **(H5.2a)** may be simply replaced by

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{\substack{i,j \\ i\neq j}} |\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}|\right]^2 = 0.$$

Therefore, **(H5.2a)** means that the influence of two Y-values on the probability of connection of two couples of random points tends to zero as $n \to \infty$.

As for assumption **(H5.2b)**, it holds whenever the correlation between the noise and the probability of connection of two couples of random points vanishes fast enough, as $n \to \infty$. Note that, in the simple case where the partition is independent of the Y_i's, the correlations in **(H5.2b)** are zero, so that **(H5.2)** is trivially satisfied. It is also verified in the noiseless case, that is, when $Y = m(\mathbf{X})$. However, in the most general context, the partitions strongly depend on the whole sample $\mathcal{D}_n$ and, unfortunately, we do not know whether **(H5.2)** is satisfied or not.

Theorem 5.2. *Assume that* **(H5.1)** *and* **(H5.2)** *are satisfied and let $t_n = a_n$. Then, provided $a_n \to \infty$ and $a_n \log n/n \to 0$, random forests are consistent, i.e.,*

$$\lim_{n\to\infty} \mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

Up to our knowledge, apart from the fact that bootstrapping is replaced by subsampling, Theorem 5.1 and Theorem 5.2 are the first consistency results for Breiman's [2001] forests. Indeed, most models studied so far are designed independently of $\mathcal{D}_n$ and are, consequently, an unrealistic representation of the true procedure. In fact, understanding Breiman's random forest behavior deserves a more involved mathematical treatment. Section 4 below offers a thorough description of the various mathematical forces in action.

Our study also sheds some interesting light on the behavior of forests when the ambient dimension p is large but the true underlying dimension of the model is small. To see how, assume that the additive model **(H5.1)** satisfies a sparsity constraint of the form

$$Y = \sum_{j=1}^{S} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

where $S < p$ represents the true, but unknown, dimension of the model. Thus, among the p original features, it is assumed that only the first (without loss of generality) S variables are

informative. Put differently, Y is assumed to be independent of the last $(p-S)$ variables. In this dimension reduction context, the ambient dimension p can be very large, but we believe that the representation is sparse, i.e., that few components of m are non-zero. As such, the value S characterizes the sparsity of the model: the smaller S, the sparser m.

Proposition 5.1 below shows that random forests nicely adapt to the sparsity setting by asymptotically performing, with high probability, splits along the S informative variables.

In this proposition, we set $m_{\text{try}} = p$ and, for all k, we denote by $j_{1,n}(\mathbf{X}), \ldots, j_{k,n}(\mathbf{X})$ the first k cut directions used to construct the cell containing $\mathbf{X}$, with the convention that $j_{q,n}(\mathbf{X}) = \infty$ if the cell has been cut strictly less than q times.

Proposition 5.1. *Assume that* **(H5.1)** *is satisfied. Let $k \in \mathbb{N}^\star$ and $\xi > 0$. Assume that there is no interval $[a, b]$ and no $j \in \{1, \ldots, S\}$ such that m_j is constant on $[a, b]$. Then, with probability $1 - \xi$, for all n large enough, we have, for all $1 \leq q \leq k$,*

$$j_{q,n}(\mathbf{X}) \in \{1, \ldots, S\}.$$

This proposition provides an interesting perspective on why random forests are still able to do a good job in a sparse framework. Since the algorithm selects splits mostly along informative variables, everything happens as if data were projected onto the vector space generated by the S informative variables. Therefore, forests are likely to only depend upon these S variables, which supports the fact that they have good performance in sparse framework.

It remains that a substantial research effort is still needed to understand the properties of forests in a high dimensional setting, when $p = p_n$ may be substantially larger than the sample size. Unfortunately, our analysis does not carry over to this context. In particular, if high-dimensionality is modelled by letting $p_n \to \infty$, then assumption **(H5.2a)** may be too restrictive since the term $(\log a_n)^{2p-2}$ will diverge at a fast rate.

5.4 Discussion

One of the main difficulties in assessing the mathematical properties of Breiman's [2001] forests is that the construction process of the individual trees strongly depends on both the X_i's and the Y_i's. For partitions that are independent of the Y_i's, consistency can be shown by relatively simple means via Stone's [1977] theorem for local averaging estimates [see also Györfi et al., 2002, Chapter 6]. However, our partitions and trees depend upon the Y-values in the data. This makes things complicated, but mathematically interesting too. Thus, logically, the proof of Theorem 5.2 starts with an adaptation of Stone's [1977] theorem tailored for random forests, whereas the proof of Theorem 5.1 is based on consistency results of data-dependent partitions developed by Nobel [1996].

Both theorems rely on Proposition 5.2 below which stresses an important feature of the random forest mechanism. It states that the variation of the regression function m within a cell of a random tree is small provided n is large enough. To this aim, we define, for any cell A, the variation of m within A as

$$\Delta(m, A) = \sup_{\mathbf{x}, \mathbf{x}' \in A} |m(\mathbf{x}) - m(\mathbf{x}')|.$$

Furthermore, we denote by $A_n(\mathbf{X}, \Theta)$ the cell of a tree built with random parameter Θ that contains the point $\mathbf{X}$.

Proposition 5.2. *Assume that* **(H5.1)** *holds. Then, for all $\rho, \xi > 0$, there exists $N \in \mathbb{N}^\star$ such that, for all $n > N$,*

$$\mathbb{P}\left[\Delta(m, A_n(\mathbf{X}, \Theta)) \leq \xi\right] \geq 1 - \rho.$$

It should be noted that in the standard, Y-independent analysis of partitioning regression function estimates, the variance is controlled by letting the diameters of the tree cells tend to zero in probability. Instead of such a geometrical assumption, Proposition 5.2 ensures that the variation of m inside a cell is small, thereby forcing the approximation error of the forest to asymptotically approach zero.

While Proposition 5.2 offers a good control of the approximation error of the forest in both regimes, a separated analysis is required for the estimation error. In regime 1 (Theorem 5.1), the parameter t_n allows to control the structure of the tree. This is in line with standard tree consistency approaches [see, e.g., Chapter 20 in Devroye et al., 1996]. Things are different for the second regime (Theorem 5.2), in which individual trees are fully grown. In that case, the estimation error is controlled by forcing the subsampling rate a_n/n to be $\mathrm{o}(1/\log n)$, which is a more unusual requirement and deserves some remarks.

At first, we note that the $\log n$ term in Theorem 5.2 is used to control the Gaussian noise ε. Thus, if the noise is assumed to be a bounded random variable, then the $\log n$ term disappears, and the condition reduces to $a_n/n \to 0$. The requirement $a_n \log n/n \to 0$ guarantees that every single observation $(\mathbf{X}_i, Y_i)$ is used in the tree construction with a probability that becomes small with n. It also implies that the query point $\mathbf{x}$ is not connected to the same data point in a high proportion of trees. If not, the predicted value at $\mathbf{x}$ would be too much influenced by one single pair $(\mathbf{X}_i, Y_i)$, making the forest inconsistent. In fact, the proof of Theorem 5.2 reveals that the estimation error of a forest estimate is small as soon as the maximum probability of connection between the query point and all observations is small. Thus, the assumption on the subsampling rate is just a convenient way to control these probabilities, by ensuring that partitions are dissimilar enough (i.e. by ensuring that $\mathbf{x}$ is connected with many data points through the forest). This idea of diversity among trees was introduced by Breiman [2001], but is generally difficult to analyse. In our approach, the subsampling is the key component for imposing tree diversity.

Theorem 5.2 comes at the price of assumption **(H5.2)**, for which we do not know if it is valid in all generality. On the other hand, Theorem 5.2, which mimics almost perfectly the algorithm used in practice, is an important step towards understanding Breiman's random forests. Contrary to most previous works, Theorem 5.2 assumes that there is only one observation per leaf of each individual tree. This implies that the single trees are eventually not consistent, since standard conditions for tree consistency require that the number of observations in the terminal nodes tends to infinity as n grows [see, e.g., Devroye et al., 1996, Györfi et al., 2002]. Thus, the random forest algorithm aggregates rough individual tree predictors to build a provably consistent general architecture.

It is also interesting to note that our results (in particular Lemma 5.3) cannot be directly extended to establish the pointwise consistency of random forests, that is, for almost all $\mathbf{x} \in$

$[0,1]^p$,

$$\lim_{n\to\infty} \mathbb{E}\big[m_{\infty,n}(\mathbf{x}) - m(\mathbf{x})\big]^2 = 0.$$

Fixing $\mathbf{x} \in [0,1]^p$, the difficulty results from the fact that we do not have a control on the diameter of the cell $A_n(\mathbf{x}, \Theta)$, whereas, since the cells form a partition of $[0,1]^p$, we have a global control on their diameters. Thus, as highlighted by Wager [2014], random forests can be inconsistent at some fixed point $\mathbf{x} \in [0,1]^p$, particularly near the edges, while being $\mathbb{L}^2$ consistent.

Let us finally mention that all results can be extended to the case where ε is a heteroscedastic and sub-Gaussian noise, with for all $\mathbf{x} \in [0,1]^p$, $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \leq \sigma'^2$, for some constant σ'^2. All proofs can be readily extended to match this context, at the price of easy technical adaptations.

5.5 Proof of Theorem 5.1 and Theorem 5.2

For the sake of clarity, proofs of the intermediary results are gathered in in the supplemental article [Scornet et al., 2015a]. We start with some notations.

5.5.1 Notations

In the sequel, to clarify the notations, we will sometimes write $d = (d^{(1)}, d^{(2)})$ to represent a cut (j, z).

Recall that, for any cell A, $\mathcal{C}_A$ is the set of all possible cuts in A. Thus, with this notation, $\mathcal{C}_{[0,1]^p}$ is just the set of all possible cuts at the root of the tree, that is, all possible choices $d = (d^{(1)}, d^{(2)})$ with $d^{(1)} \in \{1, \ldots, p\}$ and $d^{(2)} \in [0,1]$.

More generally, for any $\mathbf{x} \in [0,1]^p$, we call $\mathcal{A}_k(\mathbf{x})$ the collection of all possible $k \geq 1$ consecutive cuts used to build the cell containing $\mathbf{x}$. Such a cell is obtained after a sequence of cuts $\mathbf{d}_k = (d_1, \ldots, d_k)$, where the dependency of $\mathbf{d}_k$ upon $\mathbf{x}$ is understood. Accordingly, for any $\mathbf{d}_k \in \mathcal{A}_k(\mathbf{x})$, we let $A(\mathbf{x}, \mathbf{d}_k)$ be the cell containing $\mathbf{x}$ built with the particular k-tuple of cuts $\mathbf{d}_k$. The proximity between two elements $\mathbf{d}_k$ and $\mathbf{d}'_k$ in $\mathcal{A}_k(\mathbf{x})$ will be measured via

$$\|\mathbf{d}_k - \mathbf{d}'_k\|_\infty = \sup_{1\leq j\leq k} \max\left(|d_j^{(1)} - d_j'^{(1)}|, |d_j^{(2)} - d_j'^{(2)}|\right).$$

Accordingly, the distance d_∞ between $\mathbf{d}_k \in \mathcal{A}_k(\mathbf{x})$ and any $\mathcal{A} \subset \mathcal{A}_k(\mathbf{x})$ is

$$d_\infty(\mathbf{d}_k, \mathcal{A}) = \inf_{\mathbf{z}\in\mathcal{A}} \|\mathbf{d}_k - \mathbf{z}\|_\infty.$$

Remember that $A_n(\mathbf{X}, \Theta)$ denotes the cell of a tree containing $\mathbf{X}$ and designed with random parameter Θ. Similarly, $A_{k,n}(\mathbf{X}, \Theta)$ is the same cell but where only the first k cuts are performed ($k \in \mathbb{N}^\star$ is a parameter to be chosen later). We also denote by $\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta) = (\hat{d}_{1,n}(\mathbf{X}, \Theta), \ldots, \hat{d}_{k,n}(\mathbf{X}, \Theta))$ the k cuts used to construct the cell $A_{k,n}(\mathbf{X}, \Theta)$.

Recall that, for any cell A, the empirical criterion used to split A in the random forest algorithm is defined in (5.2). For any cut $(j, z) \in \mathcal{C}_A$, we denote the following theoretical version

of $L_n(\cdot,\cdot)$ by

$$L^\star(j,z) = \mathbb{V}[Y|\mathbf{X}\in A] - \mathbb{P}[\mathbf{X}^{(j)} < z \,|\, \mathbf{X}\in A]\ \mathbb{V}[Y|\mathbf{X}^{(j)} < z, \mathbf{X}\in A] \\ - \mathbb{P}[\mathbf{X}^{(j)} \geq z \,|\, \mathbf{X}\in A]\ \mathbb{V}[Y|\mathbf{X}^{(j)} \geq z, \mathbf{X}\in A].$$

Observe that $L^\star(\cdot,\cdot)$ does not depend upon the training set and that, by the strong law of large numbers, $L_n(j,z) \to L^\star(j,z)$ almost surely as $n\to\infty$ for all cuts $(j,z)\in\mathcal{C}_A$. Therefore, it is natural to define the best theoretical split $(j^\star,z^\star)$ of the cell A as

$$(j^\star,z^\star) \in \underset{\substack{(j,z)\in\mathcal{C}_A \\ j\in\mathcal{M}_{\text{try}}}}{\arg\min}\, L^\star(j,z).$$

In view of this criterion, we define the theoretical random forest as before, but with consecutive cuts performed by optimizing $L^\star(\cdot,\cdot)$ instead of $L_n(\cdot,\cdot)$. We note that this new forest does depend on Θ through $\mathcal{M}_{\text{try}}$, but not on the sample $\mathcal{D}_n$. In particular, the stopping criterion for dividing cells has to be changed in the theoretical random forest; instead of stopping when a cell has a single training point, we impose that each tree of the theoretical forest is stopped at a fixed level $k\in\mathbb{N}^\star$. We also let $A_k^\star(\mathbf{X},\Theta)$ be a cell of the theoretical random tree at level k, containing $\mathbf{X}$, designed with randomness Θ, and resulting from the k theoretical cuts $\mathbf{d}_k^\star(\mathbf{X},\Theta) = (d_1^\star(\mathbf{X},\Theta),\ldots,d_k^\star(\mathbf{X},\Theta))$. Since there can exist multiple best cuts at, at least, one node, we call $\mathcal{A}_k^\star(\mathbf{X},\Theta)$ the set of all k-tuples $\mathbf{d}_k^\star(\mathbf{X},\Theta)$ of best theoretical cuts used to build $A_k^\star(\mathbf{X},\Theta)$.

We are now equipped to prove Proposition 5.2. For clarity reasons, the proof has been divided in three steps. Firstly, we study in Lemma 5.1 the theoretical random forest. Then we prove in Lemma 5.3 (via Lemma 5.2), that theoretical and empirical cuts are close to each other. Proposition 5.2 is finally established as a consequence of Lemma 5.1 and Lemma 5.3. Proofs of these lemmas are to be found in the supplemental article [Scornet et al., 2015a].

5.5.2 Proof of Proposition 5.2

We first need a lemma which states that the variation of $m(\mathbf{X})$ within the cell $A_k^\star(\mathbf{X},\Theta)$ where $\mathbf{X}$ falls, as measured by $\Delta(m, A_k^\star(\mathbf{X},\Theta))$, tends to zero.

Lemma 5.1. *Assume that* **(H5.1)** *is satisfied. Then, for all* $\mathbf{x}\in[0,1]^p$,

$$\Delta(m, A_k^\star(\mathbf{x},\Theta)) \to 0, \quad \textit{almost surely, as } k\to\infty.$$

The next step is to show that cuts in theoretical and original forests are close to each other. To this aim, for any $\mathbf{x}\in[0,1]^p$ and any k-tuple of cuts $\mathbf{d}_k\in\mathcal{A}_k(\mathbf{x})$, we define

$$L_{n,k}(\mathbf{x},\mathbf{d}_k) = \frac{1}{N_n(A(\mathbf{x},\mathbf{d}_{k-1}))}\sum_{i=1}^n (Y_i - \bar{Y}_{A(\mathbf{x},\mathbf{d}_{k-1})})^2 \mathbb{1}_{\mathbf{X}_i\in A(\mathbf{x},\mathbf{d}_{k-1})} \\ - \frac{1}{N_n(A(\mathbf{x},\mathbf{d}_{k-1}))}\sum_{i=1}^n \Big(Y_i - \bar{Y}_{A_L(\mathbf{x},\mathbf{d}_{k-1})}\mathbb{1}_{\mathbf{X}_i^{(d_k^{(1)})} < d_k^{(2)}} \\ - \bar{Y}_{A_R(\mathbf{x},\mathbf{d}_{k-1})}\mathbb{1}_{\mathbf{X}_i^{(d_k^{(1)})} \geq d_k^{(2)}}\Big)^2 \mathbb{1}_{\mathbf{X}_i\in A(\mathbf{x},\mathbf{d}_{k-1})},$$

where $A_L(\mathbf{x}, \mathbf{d}_{k-1}) = A(\mathbf{x}, \mathbf{d}_{k-1}) \cap \{\mathbf{z} : \mathbf{z}^{(d_k^{(1)})} < d_k^{(2)}\}$ and $A_R(\mathbf{x}, \mathbf{d}_{k-1}) = A(\mathbf{x}, \mathbf{d}_{k-1}) \cap \{\mathbf{z} : \mathbf{z}^{(d_k^{(1)})} \geq d_k^{(2)}\}$, and where we use the convention $0/0 = 0$ when $A(\mathbf{x}, \mathbf{d}_{k-1})$ is empty. Besides, we let $A(\mathbf{x}, \mathbf{d}_0) = [0,1]^p$ in the previous equation. The quantity $L_{n,k}(\mathbf{x}, \mathbf{d}_k)$ is nothing but the criterion to maximize in d_k to find the best k-th cut in the cell $A(\mathbf{x}, \mathbf{d}_{k-1})$. Lemma 5.2 below ensures that $L_{n,k}(\mathbf{x}, \cdot)$ is stochastically equicontinuous, for all $\mathbf{x} \in [0,1]^p$. To this aim, for all $\xi > 0$, and for all $\mathbf{x} \in [0,1]^p$, we denote by $\mathcal{A}_{k-1}^{\xi}(\mathbf{x}) \subset \mathcal{A}_{k-1}(\mathbf{x})$ the set of all $(k-1)$-tuples $\mathbf{d}_{k-1}$ such that the cell $A(\mathbf{x}, \mathbf{d}_{k-1})$ contains a hypercube of edge length ξ. Moreover, we let $\bar{\mathcal{A}}_k^{\xi}(\mathbf{x}) = \{\mathbf{d}_k : \mathbf{d}_{k-1} \in \mathcal{A}_{k-1}^{\xi}(\mathbf{x})\}$ equipped with the norm $\|\mathbf{d}_k\|_\infty$.

Lemma 5.2. *Assume that* **(H5.1)** *is satisfied. Fix* $\mathbf{x} \in [0,1]^p$, $k \in \mathbb{N}^\star$, *and let* $\xi > 0$. *Then* $L_{n,k}(\mathbf{x}, \cdot)$ *is stochastically equicontinuous on* $\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})$, *that is, for all* $\alpha, \rho > 0$, *there exists* $\delta > 0$ *such that*

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{\substack{\|\mathbf{d}_k - \mathbf{d}'_k\|_\infty \leq \delta \\ \mathbf{d}_k, \mathbf{d}'_k \in \bar{\mathcal{A}}_k^{\xi}(\mathbf{x})}} |L_{n,k}(\mathbf{x}, \mathbf{d}_k) - L_{n,k}(\mathbf{x}, \mathbf{d}'_k)| > \alpha\right] \leq \rho.$$

Lemma 5.2 is then used in Lemma 5.3 to assess the distance between theoretical and empirical cuts.

Lemma 5.3. *Assume that* **(H5.1)** *is satisfied. Fix* $\xi, \rho > 0$ *and* $k \in \mathbb{N}^\star$. *Then there exists* $N \in \mathbb{N}^\star$ *such that, for all* $n \geq N$,

$$\mathbb{P}\left[d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \leq \xi\right] \geq 1 - \rho.$$

We are now ready to prove Proposition 5.2. Fix $\rho, \xi > 0$. Since almost sure convergence implies convergence in probability, according to Lemma 5.1, there exists $k_0 \in \mathbb{N}^\star$ such that

$$\mathbb{P}\left[\Delta(m, A_{k_0}^\star(\mathbf{X}, \Theta)) \leq \xi\right] \geq 1 - \rho. \tag{5.3}$$

By Lemma 5.3, for all $\xi_1 > 0$, there exists $N \in \mathbb{N}^\star$ such that, for all $n \geq N$,

$$\mathbb{P}\left[d_\infty(\hat{\mathbf{d}}_{k_0,n}(\mathbf{X}, \Theta), \mathcal{A}_{k_0}^\star(\mathbf{X}, \Theta)) \leq \xi_1\right] \geq 1 - \rho. \tag{5.4}$$

Since m is uniformly continuous, we can choose ξ_1 sufficiently small such that, for all $\mathbf{x} \in [0,1]^p$, for all $\mathbf{d}_{k_0}, \mathbf{d}'_{k_0}$ satisfying $d_\infty(\mathbf{d}_{k_0}, \mathbf{d}'_{k_0}) \leq \xi_1$, we have

$$\left|\Delta(m, A(\mathbf{x}, \mathbf{d}_{k_0})) - \Delta(m, A(\mathbf{x}, \mathbf{d}'_{k_0}))\right| \leq \xi. \tag{5.5}$$

Thus, combining inequalities (5.4) and (5.5), we obtain

$$\mathbb{P}\Big[\left|\Delta(m, A_{k_0,n}(\mathbf{X}, \Theta)) - \Delta(m, A_{k_0}^\star(\mathbf{X}, \Theta))\right| \leq \xi\Big] \geq 1 - \rho. \tag{5.6}$$

Using the fact that $\Delta(m, A) \leq \Delta(m, A')$ whenever $A \subset A'$, we deduce from (5.3) and (5.6) that, for all $n \geq N$,

$$\mathbb{P}\left[\Delta(m, A_n(\mathbf{X}, \Theta)) \leq 2\xi\right] \geq 1 - 2\rho.$$

This concludes the proof of Proposition 5.2.

5.5.3 Proof of Theorem 5.1

We still need some additional notations. The partition obtained with the random variable Θ and the data set $\mathcal{D}_n$ is denoted by $\mathcal{P}_n(\mathcal{D}_n, \Theta)$, which we abbreviate as $\mathcal{P}_n(\Theta)$. We let

$$\Pi_n(\Theta) = \{\mathcal{P}((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), \Theta) : (\mathbf{x}_i, y_i) \in [0,1]^p \times \mathbb{R}\}$$

be the family of all achievable partitions with random parameter Θ. Accordingly, we let

$$M(\Pi_n(\Theta)) = \max\{\text{Card}(\mathcal{P}) : \mathcal{P} \in \Pi_n(\Theta)\}$$

be the maximal number of terminal nodes among all partitions in $\Pi_n(\Theta)$. Given a set $\mathbf{z}_1^n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \subset [0,1]^p$, $\Gamma(\mathbf{z}_1^n, \Pi_n(\Theta))$ denotes the number of distinct partitions of $\mathbf{z}_1^n$ induced by elements of $\Pi_n(\Theta)$, that is, the number of different partitions $\{\mathbf{z}_1^n \cap A : A \in \mathcal{P}\}$ of $\mathbf{z}_1^n$, for $\mathcal{P} \in \Pi_n(\Theta)$. Consequently, the partitioning number $\Gamma_n(\Pi_n(\Theta))$ is defined by

$$\Gamma_n(\Pi_n(\Theta)) = \max\{\Gamma(\mathbf{z}_1^n, \Pi_n(\Theta)) : \mathbf{z}_1, \ldots, \mathbf{z}_n \in [0,1]^p\}.$$

Let $(\beta_n)_n$ be a positive sequence, and define the truncated operator T_{β_n} by

$$\begin{cases} T_{\beta_n} u = u & \text{if } |u| < \beta_n \\ T_{\beta_n} u = \text{sign}(u)\beta_n & \text{if } |u| \geq \beta_n. \end{cases}$$

Hence, $T_{\beta_n} m_n(\mathbf{X}, \Theta)$, $Y_L = T_L Y$ and $Y_{i,L} = T_L Y_i$ are defined unambiguously. We let $\mathcal{F}_n(\Theta)$ be the set of all functions $f : [0,1]^p \to \mathbb{R}$ piecewise constant on each cell of the partition $\mathcal{P}_n(\Theta)$. (Notice that $\mathcal{F}_n(\Theta)$ depends on the whole data set.) Finally, we denote by $\mathcal{I}_{n,\Theta}$ the set of indices of the data points that are selected during the subsampling step. Thus the tree estimate $m_n(\mathbf{x}, \Theta)$ satisfies

$$m_n(\cdot, \Theta) \in \underset{f \in \mathcal{F}_n(\Theta)}{\text{argmin}} \frac{1}{a_n} \sum_{i \in \mathcal{I}_{n,\Theta}} |f(\mathbf{X}_i) - Y_i|^2.$$

The proof of Theorem 5.1 is based on ideas developed by Nobel [1996], and worked out in Theorem 10.2 in Györfi et al. [2002]. This theorem, tailored for our context, is recalled below for the sake of completeness.

Theorem 5.3. *[Györfi et al., 2002] Let m_n and $\mathcal{F}_n(\Theta)$ be as above. Assume that*

(i) $\lim_{n\to\infty} \beta_n = \infty$,

(ii) $$\lim_{n\to\infty} \mathbb{E}\left[\inf_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \leq \beta_n}} \mathbb{E}_{\mathbf{X}} \left[f(\mathbf{X}) - m(\mathbf{X})\right]^2 \right] = 0,$$

(iii) For all $L > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \leq \beta_n}} \left| \frac{1}{a_n} \sum_{i \in \mathcal{I}_{n,\Theta}} \left[f(\mathbf{X}_i) - Y_{i,L}\right]^2 - \mathbb{E}\left[f(\mathbf{X}) - Y_L\right]^2 \right| \right] = 0.$$

Then

$$\lim_{n\to\infty} \mathbb{E}\left[T_{\beta_n} m_n(\mathbf{X}, \Theta) - m(\mathbf{X})\right]^2 = 0.$$

Statement (ii) (resp. statement (iii)) allows us to control the approximation error (resp. the estimation error) of the truncated estimate. Since the truncated estimate $T_{\beta_n} m_n$ is piecewise constant on each cell of the partition $\mathcal{P}_n(\Theta)$, $T_{\beta_n} m_n$ belongs to the set $\mathcal{F}_n(\Theta)$. Thus, the term in (ii) is the classical approximation error.

We are now equipped to prove Theorem 5.1. Fix $\xi > 0$ and note that we just have to check statements $(i) - (iii)$ of Theorem 5.3 to prove that the truncated estimate of the random forest is consistent. Throughout the proof, we let $\beta_n = \|m\|_\infty + \sigma\sqrt{2}(\log a_n)^2$. Clearly, statement (i) is true.

Approximation error To prove (ii), let

$$f_{n,\Theta} = \sum_{A \in \mathcal{P}_n(\Theta)} m(\mathbf{z}_A) \mathbb{1}_A,$$

where $\mathbf{z}_A \in A$ is an arbitrary point picked in cell A. Since, according to **(H5.1)**, $\|m\|_\infty < \infty$, for all n large enough such that $\beta_n > \|m\|_\infty$, we have

$$\begin{aligned}
\mathbb{E} \inf_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \le \beta_n}} \mathbb{E}_{\mathbf{X}} \left[f(\mathbf{X}) - m(\mathbf{X})\right]^2 &\le \mathbb{E} \inf_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \le \|m\|_\infty}} \mathbb{E}_{\mathbf{X}} \left[f(\mathbf{X}) - m(\mathbf{X})\right]^2 \\
&\le \mathbb{E}\big[f_{\Theta,n}(\mathbf{X}) - m(\mathbf{X})\big]^2 \\
&\quad \text{(since } f_{\Theta,n} \in \mathcal{F}_n(\Theta)\text{)} \\
&\le \mathbb{E}\big[m(\mathbf{z}_{A_n(\mathbf{X},\Theta)}) - m(\mathbf{X})\big]^2 \\
&\le \mathbb{E}\big[\Delta(m, A_n(\mathbf{X}, \Theta))\big]^2 \\
&\le \xi^2 + 4\|m\|_\infty^2 \mathbb{P}\big[\Delta(m, A_n(\mathbf{X}, \Theta)) > \xi\big].
\end{aligned}$$

Thus, using Proposition 5.2, we see that, for all n large enough,

$$\mathbb{E} \inf_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \le \beta_n}} \mathbb{E}_{\mathbf{X}} \left[f(\mathbf{X}) - m(\mathbf{X})\right]^2 \le 2\xi^2.$$

This establishes (ii).

Estimation error To prove statement (iii), fix $L > 0$. Then, for all n large enough such that $L < \beta_n$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X},\mathcal{D}_n}\left(\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\left|\frac{1}{a_n}\sum_{i\in\mathcal{I}_{n,\Theta}}\big[f(\mathbf{X}_i)-Y_{i,L}\big]^2-\mathbb{E}\big[f(\mathbf{X})-Y_L\big]^2\right|>\xi\right)\\
&\le 8\exp\left[\log\Gamma_n(\Pi_n(\Theta))+2M(\Pi_n(\Theta))\log\left(\frac{333e\beta_n^2}{\xi}\right)-\frac{a_n\xi^2}{2048\beta_n^4}\right]\\
&\quad\text{[according to Theorem 9.1 in Györfi et al., 2002]}\\
&\le 8\exp\left[-\frac{a_n}{\beta_n^4}\left(\frac{\xi^2}{2048}-\frac{\beta_n^4\log\Gamma_n(\Pi_n)}{a_n}-\frac{2\beta_n^4M(\Pi_n)}{a_n}\log\left(\frac{333e\beta_n^2}{\xi}\right)\right)\right].
\end{aligned}$$

Since each tree has exactly t_n terminal nodes, we have $M(\Pi_n(\Theta)) = t_n$ and simple calculations show that

$$\Gamma_n(\Pi_n(\Theta)) \le (da_n)^{t_n}.$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\left|\frac{1}{a_n}\sum_{i\in\mathcal{I}_{n,\Theta}}\big[f(\mathbf{X}_i)-Y_{i,L}\big]^2-\mathbb{E}\big[f(\mathbf{X})-Y_L\big]^2\right|>\xi\right)\\
&\le 8\exp\left(-\frac{a_nC_{\xi,n}}{\beta_n^4}\right),
\end{aligned}$$

where

$$\begin{aligned}
C_{\xi,n} &= \frac{\xi^2}{2048}-4\sigma^4\frac{t_n(\log(da_n))^9}{a_n}-8\sigma^4\frac{t_n(\log a_n)^8}{a_n}\log\left(\frac{666e\sigma^2(\log a_n)^4}{\xi}\right)\\
&\to \frac{\xi^2}{2048}, \quad \text{as } n\to\infty,
\end{aligned}$$

by our assumption. Finally, observe that

$$\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\left|\frac{1}{a_n}\sum_{i\in\mathcal{I}_{n,\Theta}}\big[f(\mathbf{X}_i)-Y_{i,L}\big]^2-\mathbb{E}\big[f(\mathbf{X})-Y_L\big]^2\right| \le 2(\beta_n+L)^2,$$

which yields, for all n large enough,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\Big|\frac{1}{a_n}\sum_{i=1}^{a_n}\big[f(\mathbf{X}_i)-Y_{i,L}\big]^2-\mathbb{E}\big[f(\mathbf{X})-Y_L\big]^2\Big|\Big]\le\xi\\
&\quad+2(\beta_n+L)^2\mathbb{P}\Big[\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\Big|\frac{1}{a_n}\sum_{i=1}^{a_n}\big[f(\mathbf{X}_i)-Y_{i,L}\big]^2-\mathbb{E}\big[f(\mathbf{X})-Y_L\big]^2\Big|>\xi\Big]\\
&\le\xi+16(\beta_n+L)^2\exp\left(-\frac{a_nC_{\xi,n}}{\beta_n^4}\right)\\
&\le 2\xi.
\end{aligned}$$

Thus, according to Theorem 5.3,

$$\mathbb{E}\big[T_{\beta_n}m_n(\mathbf{X},\Theta)-m(\mathbf{X})\big]^2\to 0.$$

Untruncated estimate It remains to show the consistency of the non truncated random forest estimate, and the proof will be complete. For that purpose, note that, for all n large enough,

$$\begin{aligned}
\mathbb{E}\big[m_{\infty,n}(\mathbf{X})-m(\mathbf{X})\big]^2&=\mathbb{E}\big[\mathbb{E}_\Theta[m_n(\mathbf{X},\Theta)]-m(\mathbf{X})\big]^2\\
&\le\mathbb{E}\big[m_n(\mathbf{X},\Theta)-m(\mathbf{X})\big]^2\\
&\quad\text{(by Jensen's inequality)}\\
&\le\mathbb{E}\big[m_n(\mathbf{X},\Theta)-T_{\beta_n}m_n(\mathbf{X},\Theta)\big]^2\\
&\qquad+\mathbb{E}\big[T_{\beta_n}m_n(\mathbf{X},\Theta)-m(\mathbf{X})\big]^2\\
&\le\mathbb{E}\Big[\big[m_n(\mathbf{X},\Theta)-T_{\beta_n}m_n(\mathbf{X},\Theta)\big]^2\mathbb{1}_{m_n(\mathbf{X},\Theta)\ge\beta_n}\Big]+\xi\\
&\le\mathbb{E}\Big[m_n^2(\mathbf{X},\Theta)\mathbb{1}_{m_n(\mathbf{X},\Theta)\ge\beta_n}\Big]+\xi\\
&\le\mathbb{E}\Big[\mathbb{E}\Big[m_n^2(\mathbf{X},\Theta)\mathbb{1}_{m_n(\mathbf{X},\Theta)\ge\beta_n}|\Theta\Big]\Big]+\xi.
\end{aligned}$$

Since $|m_n(\mathbf{X},\Theta)|\le\|m\|_\infty+\max\limits_{1\le i\le n}|\varepsilon_i|$, we have

$$\begin{aligned}
&\mathbb{E}\Big[m_n^2(\mathbf{X},\Theta)\mathbb{1}_{m_n(\mathbf{X},\Theta)\ge\beta_n}|\Theta\Big]\\
&\le\mathbb{E}\Big[(2\|m\|_\infty^2+2\max_{1\le i\le a_n}\varepsilon_i^2)\mathbb{1}_{\max\limits_{1\le i\le a_n}\varepsilon_i\ge\sigma\sqrt{2}(\log a_n)^2}\Big]\\
&\le 2\|m\|_\infty^2\mathbb{P}\big[\max_{1\le i\le a_n}\varepsilon_i\ge\sigma\sqrt{2}(\log a_n)^2\big]\\
&\qquad+2\Big(\mathbb{E}\big[\max_{1\le i\le a_n}\varepsilon_i^4\big]\mathbb{P}\big[\max_{1\le i\le a_n}\varepsilon_i\ge\sigma\sqrt{2}(\log a_n)^2\big]\Big)^{1/2}.
\end{aligned}$$

It is easy to see that

$$\mathbb{P}\big[\max_{1\le i\le a_n}\varepsilon_i\ge\sigma\sqrt{2}(\log a_n)^2\big]\le\frac{a_n^{1-\log a_n}}{2\sqrt{\pi}(\log a_n)^2}.$$

Finally, since the ε_i's are centered i.i.d. Gaussian random variables, we have, for all n large enough,

$$\begin{aligned}\mathbb{E}\big[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\big]^2 &\leq \frac{2\|m\|_\infty^2 a_n^{1-\log a_n}}{2\sqrt{\pi}(\log a_n)^2} + \xi + 2\left(3a_n\sigma^4\frac{a_n^{1-\log a_n}}{2\sqrt{\pi}(\log a_n)^2}\right)^{1/2} \\ &\leq 3\xi.\end{aligned}$$

This concludes the proof of Theorem 5.1.

5.5.4 Proof of Theorem 5.2

Recall that each cell contains exactly one data point. Thus, letting

$$W_{ni}(\mathbf{X}) = \mathbb{E}_\Theta\left[\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}\right],$$

the random forest estimate $m_{\infty,n}$ may be rewritten as

$$m_{\infty,n}(\mathbf{X}) = \sum_{i=1}^n W_{ni}(\mathbf{X})Y_i.$$

We have in particular that $\sum_{i=1}^n W_{ni}(\mathbf{X}) = 1$. Thus,

$$\begin{aligned}\mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 &\leq 2\mathbb{E}\left[\sum_{i=1}^n W_{ni}(\mathbf{X})(Y_i - m(\mathbf{X}_i))\right]^2 \\ &\quad + 2\mathbb{E}\left[\sum_{i=1}^n W_{ni}(\mathbf{X})(m(\mathbf{X}_i) - m(\mathbf{X}))\right]^2 \\ &\stackrel{\text{def}}{=} 2I_n + 2J_n.\end{aligned}$$

Approximation error Fix $\alpha > 0$. To upper bound J_n, note that by Jensen's inequality,

$$\begin{aligned}J_n &\leq \mathbb{E}\left[\sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}(m(\mathbf{X}_i) - m(\mathbf{X}))^2\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}\Delta^2(m, A_n(\mathbf{X},\Theta))\right] \\ &\leq \mathbb{E}\left[\Delta^2(m, A_n(\mathbf{X},\Theta))\right].\end{aligned}$$

So, by definition of $\Delta(m, A_n(\mathbf{X},\Theta))^2$,

$$\begin{aligned}J_n &\leq 4\|m\|_\infty^2\,\mathbb{E}[\mathbb{1}_{\Delta^2(m,A_n(\mathbf{X},\Theta))\geq\alpha}] + \alpha \\ &\leq \alpha(4\|m\|_\infty^2 + 1),\end{aligned}$$

for all n large enough, according to Proposition 5.2.

Estimation error To bound I_n from above, we note that

$$\begin{aligned} I_n &= \mathbb{E}\left[\sum_{i,j=1}^{n} W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right] \\ &= \mathbb{E}\left[\sum_{i=1} W_{ni}^2(\mathbf{X})(Y_i - m(\mathbf{X}_i))^2\right] + I_n', \end{aligned}$$

where

$$I_n' = \mathbb{E}\left[\sum_{\substack{i,j \\ i\neq j}} \mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\mathbf{X}\overset{\Theta'}{\leftrightarrow}\mathbf{X}_j}(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right].$$

The term I_n', which involves the double products, is handled separately in Lemma 5.4 below. According to this lemma, and by assumption **(H5.2)**, for all n large enough,

$$|I_n'| \leq \alpha.$$

Consequently, recalling that $\varepsilon_i = Y_i - m(\mathbf{X}_i)$, we have, for all n large enough,

$$\begin{aligned} |I_n| &\leq \alpha + \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})(Y_i - m(\mathbf{X}_i))^2\right] \\ &\leq \alpha + \mathbb{E}\left[\max_{1\leq\ell\leq n} W_{n\ell}(\mathbf{X}) \sum_{i=1}^{n} W_{ni}(\mathbf{X})\varepsilon_i^2\right] \\ &\leq \alpha + \mathbb{E}\left[\max_{1\leq\ell\leq n} W_{n\ell}(\mathbf{X}) \max_{1\leq i\leq n} \varepsilon_i^2\right]. \end{aligned} \tag{5.7}$$

Now, observe that in the subsampling step, there are exactly $\binom{a_n-1}{n-1}$ choices to pick a fixed observation $\mathbf{X}_i$. Since $\mathbf{x}$ and $\mathbf{X}_i$ belong to the same cell only if $\mathbf{X}_i$ is selected in the subsampling step, we see that

$$\mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right] \leq \frac{\binom{a_n-1}{n-1}}{\binom{a_n}{n}} = \frac{a_n}{n},$$

where $\mathbb{P}_\Theta$ denotes the probability with respect to Θ, conditional on $\mathbf{X}$ and $\mathcal{D}_n$. So,

$$\max_{1\leq i\leq n} W_{ni}(\mathbf{X}) \leq \max_{1\leq i\leq n} \mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right] \leq \frac{a_n}{n}. \tag{5.8}$$

Thus, combining inequalities (5.7) and (5.8), for all n large enough,

$$|I_n| \leq \alpha + \frac{a_n}{n}\,\mathbb{E}\left[\max_{1\leq i\leq n} \varepsilon_i^2\right].$$

The term inside the brackets is the maximum of n χ^2-squared distributed random variables. Thus, for some positive constant C,

$$\mathbb{E}\left[\max_{1\leq i\leq n} \varepsilon_i^2\right] \leq C \log n$$

[see, e.g., Chapter 1 in Boucheron et al., 2013]. We conclude that, for all n large enough,

$$I_n \leq \alpha + C\,\frac{a_n \log n}{n} \leq 2\alpha.$$

Since α was arbitrary, the proof is complete.

Lemma 5.4. *Assume that* **(H5.2)** *is satisfied. Then, for all $\varepsilon > 0$, and all n large enough, $|I_n'| \leq \alpha$.*

Proof of Lemma 5.4. Firstly, assume that **(H5.2b)** is verified. Thus, we have for all $\ell_1, \ell_2 \in \{0,1\}$,

$$\begin{aligned}
&\operatorname{Corr}(Y_i - m(\mathbf{X}_i), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)} | \mathbf{X}_i, \mathbf{X}_j, Y_j) \\
&= \frac{\mathbb{E}\big[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}\big]}{\mathbb{V}^{1/2}\big[Y_i - m(\mathbf{X}_i)|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]\mathbb{V}^{1/2}\big[\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]} \\
&= \frac{\mathbb{E}\big[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]}{\sigma\big(\mathbb{P}\big[Z_{i,j} = (\ell_1,\ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j\big] - \mathbb{P}\big[Z_{i,j} = (\ell_1,\ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]^2\big)^{1/2}} \\
&\geq \frac{\mathbb{E}\big[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]}{\sigma\mathbb{P}^{1/2}\big[Z_{i,j} = (\ell_1,\ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]},
\end{aligned}$$

where the first equality comes from the fact that, for all $\ell_1, \ell_2 \in \{0,1\}$,

$$\begin{aligned}
&\operatorname{Cov}(Y_i - m(\mathbf{X}_i), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j) \\
&\quad = \mathbb{E}\big[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j\big],
\end{aligned}$$

since $\mathbb{E}[Y_i - m(\mathbf{X}_i)|\mathbf{X}_i, \mathbf{X}_j, Y_j] = 0$. Thus, noticing that, almost surely,

$$\begin{aligned}
&\mathbb{E}\left[Y_i - m(\mathbf{X}_i)\middle| Z_{i,j}, \mathbf{X}_i, \mathbf{X}_j, Y_j\right] \\
&\quad = \sum_{\ell_1,\ell_2=1}^{2} \frac{\mathbb{E}\left[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j\right]}{\mathbb{P}\big[Z_{i,j} = (\ell_1,\ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]}\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)} \\
&\quad \leq 4\sigma \max_{\ell_1,\ell_2=0,1} \frac{|\operatorname{Corr}(Y_i - m(\mathbf{X}_i), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j)|}{\mathbb{P}^{1/2}\big[Z_{i,j} = (\ell_1,\ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j\big]} \\
&\quad \leq 4\sigma\gamma_n,
\end{aligned}$$

we conclude that the first statement in **(H5.2b)** implies that, almost surely,

$$\mathbb{E}\left[Y_i - m(\mathbf{X}_i)\middle| Z_{i,j}, \mathbf{X}_i, \mathbf{X}_j, Y_j\right] \leq 4\sigma\gamma_n.$$

Similarly, one can prove that second statement in assumption **(H5.2b)** implies that, almost surely,

$$\mathbb{E}\left[|Y_i - m(\mathbf{X}_i)|^2 \middle| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\right] \leq 4C\sigma^2.$$

Returning to the term I'_n, and recalling that $W_{ni}(\mathbf{X}) = \mathbb{E}_\Theta[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}]$, we obtain

$$\begin{aligned}
I'_n &= \mathbb{E}\left[\sum_{\substack{i,j \\ i \neq j}} \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right] \\
&= \sum_{\substack{i,j \\ i \neq j}} \mathbb{E}\Bigg[\mathbb{E}\Bigg[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j)) \\
&\qquad\qquad \Bigg| \mathbf{X}_i, \mathbf{X}_j, Y_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}\Bigg]\Bigg] \\
&= \sum_{\substack{i,j \\ i \neq j}} \mathbb{E}\Bigg[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} (Y_i - m(\mathbf{X}_i)) \\
&\qquad\qquad \times \mathbb{E}\left[Y_j - m(\mathbf{X}_j) | \mathbf{X}_i, \mathbf{X}_j, Y_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}\right]\Bigg].
\end{aligned}$$

Therefore, by assumption **(H5.2b)**,

$$\begin{aligned}
|I'_n| &\leq 4\sigma\gamma_n \sum_{\substack{i,j \\ i \neq j}} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} |Y_i - m(\mathbf{X}_i)|\right] \\
&\leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} |Y_i - m(\mathbf{X}_i)|\right] \\
&\leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{E}\left[|Y_i - m(\mathbf{X}_i)| \middle| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\right]\right] \\
&\leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{E}^{1/2}\left[|Y_i - m(\mathbf{X}_i)|^2 \middle| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\right]\right] \\
&\leq 2\sigma C^{1/2} \gamma_n.
\end{aligned}$$

This proves the result, provided **(H5.2b)** is true. Let us now assume that **(H5.2a)** is verified. The key argument is to note that a data point $\mathbf{X}_i$ can be connected with a random point $\mathbf{X}$ if $(\mathbf{X}_i, Y_i)$ is selected via the subsampling procedure and if there is no other data points in the hyperrectangle defined by $\mathbf{X}_i$ and $\mathbf{X}$. Data points $\mathbf{X}_i$ satisfying the latter geometrical property are called Layered Nearest Neighbor [LNN, see, e.g., Barndorff-Nielsen and Sobel, 1966]. The connection between LNN and random forests has been first observed by Lin and Jeon [2006],

and latter worked out by Biau and Devroye [2010]. It is known, in particular, that the number of LNN $L_{a_n}(\mathbf{X})$ among a_n data points uniformly distributed on $[0,1]^p$ satisfies, for some constant $C_1 > 0$ and for all n large enough,

$$\begin{aligned}\mathbb{E}[L_{a_n}^4(\mathbf{X})] &\leq a_n \mathbb{P}[\mathbf{X} \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}_j] + 16 a_n^2 \mathbb{P}[\mathbf{X} \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}_i] \mathbb{P}[\mathbf{X} \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}_j] \\ &\leq C_1 (\log a_n)^{2d-2}, \end{aligned} \tag{5.9}$$

[see, e.g., Barndorff-Nielsen and Sobel, 1966, Bai et al., 2005]. Thus, we have

$$I_n' = \mathbb{E}\left[\sum_{\substack{i,j \\ i \neq j}} \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j)) \right].$$

Consequently,

$$\begin{aligned} I_n' = \mathbb{E}\Bigg[\sum_{\substack{i,j \\ i \neq j}} &(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j)) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \\ &\times \mathbb{E}\Big[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \big| \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i, Y_j \Big] \Bigg], \end{aligned}$$

where $\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}$ is the event where $\mathbf{X}_i$ is selected by the subsampling and is also a LNN of $\mathbf{X}$. Next, with notations of assumption **(H5.2)**,

$$\begin{aligned} I_n' &= \mathbb{E}\Bigg[\sum_{\substack{i,j \\ i \neq j}} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j)) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \\ &\qquad \times \psi_{i,j}(Y_i, Y_j) \Bigg] \\ &= \mathbb{E}\left[\sum_{\substack{i,j \\ i \neq j}} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j)) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \psi_{i,j} \right] \\ &\quad + \mathbb{E}\Bigg[\sum_{\substack{i,j \\ i \neq j}} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j)) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \\ &\qquad \times (\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}) \Bigg]. \end{aligned}$$

The first term is easily seen to be zero since

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{\substack{i,j\\ i\neq j}}(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}}\mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}}\psi(\mathbf{X},\Theta,\Theta',\mathbf{X}_1,\ldots,\mathbf{X}_n)\Bigg]\\
&= \sum_{\substack{i,j\\ i\neq j}}\mathbb{E}\Bigg[\mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}}\mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}}\psi_{i,j}\\
&\qquad\qquad \times \mathbb{E}\big[(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\big|\mathbf{X},\mathbf{X}_1,\ldots,\mathbf{X}_n,\Theta,\Theta'\big]\Bigg]\\
&= 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|I_n'| &\leq \mathbb{E}\Bigg[\sum_{\substack{i,j\\ i\neq j}}|Y_i - m(\mathbf{X}_i)||Y_j - m(\mathbf{X}_j)|\mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}}\mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}}\\
&\qquad\qquad \times |\psi_{i,j}(Y_i,Y_j) - \psi_{i,j}|\Bigg]\\
&\leq \mathbb{E}\Bigg[\max_{1\leq\ell\leq n}|Y_i - m(\mathbf{X}_i)|^2 \max_{\substack{i,j\\ i\neq j}}|\psi_{i,j}(Y_i,Y_j) - \psi_{i,j}|\\
&\qquad\qquad \times \sum_{\substack{i,j\\ i\neq j}}\mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}}\mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}}\Bigg].
\end{aligned}$$

Now, observe that

$$\sum_{\substack{i,j\\ i\neq j}}\mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}}\mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \leq L_{a_n}^2(\mathbf{X}),$$

Consequently,

$$\begin{aligned}
|I_n'| &\leq \mathbb{E}^{1/2}\Bigg[L_{a_n}^4(\mathbf{X})\max_{1\leq\ell\leq n}|Y_i - m(\mathbf{X}_i)|^4\Bigg]\\
&\qquad\qquad \times \mathbb{E}^{1/2}\Bigg[\max_{\substack{i,j\\ i\neq j}}|\psi_{i,j}(Y_i,Y_j) - \psi_{i,j}|\Bigg]^2. \qquad (5.10)
\end{aligned}$$

Simple calculations reveal that there exists $C_1 > 0$ such that, for all n,

$$\mathbb{E}\big[\max_{1\leq\ell\leq n}|Y_i - m(\mathbf{X}_i)|^4\big] \leq C_1(\log n)^2. \qquad (5.11)$$

Thus, by inequalities (5.9) and (5.11), the first term in (5.10) can be upper bounded as follows:

$$\begin{aligned}
&\mathbb{E}^{1/2}\left[L_{a_n}^4(\mathbf{X}) \max_{1\le\ell\le n} |Y_i - m(\mathbf{X}_i)|^4\right] \\
&= \mathbb{E}^{1/2}\left[L_{a_n}^4(\mathbf{X})\mathbb{E}\big[\max_{1\le\ell\le n} |Y_i - m(\mathbf{X}_i)|^4 \big| \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n\big]\right] \\
&\le C'(\log n)(\log a_n)^{d-1}.
\end{aligned}$$

Finally,

$$|I_n'| \le C'(\log a_n)^{d-1}(\log n)^{\alpha/2}\mathbb{E}^{1/2}\left[\max_{\substack{i,j \\ i\neq j}} |\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}|\right]^2,$$

which tends to zero by assumption. □

5.6 Technical results

5.6.1 Proof of Lemma 1

Technical Lemma 5.1. *Assume that* **(H5.1)** *is satisfied and that $L^\star \equiv 0$ for all cuts in some given cell A. Then the regression function m is constant on A.*

Proof of Technical Lemma 5.1. We start by proving the result in dimension $p = 1$. Letting $A = [a, b]$ $(0 \le a < b \le 1)$, and recalling that $Y = m(\mathbf{X}) + \varepsilon$, one has

$$\begin{aligned}
L^\star(1, z) &= \mathbb{V}\left[Y | \mathbf{X} \in A\right] - \mathbb{P}\left[a \le \mathbf{X} \le z \,|\, \mathbf{X} \in A\right] \mathbb{V}\left[Y | a \le \mathbf{X} \le z\right] \\
&\quad - \mathbb{P}\left[z \le \mathbf{X} \le b \,|\, \mathbf{X} \in A\right] \mathbb{V}\left[Y | z < \mathbf{X} \le b\right] \\
&= -\frac{1}{(b-a)^2}\left(\int_a^b m(t)\mathrm{d}t\right)^2 + \frac{1}{(b-a)(z-a)}\left(\int_a^z m(t)\mathrm{d}t\right)^2 \\
&\quad + \frac{1}{(b-a)(b-z)}\left(\int_z^b m(t)\mathrm{d}t\right)^2.
\end{aligned}$$

Let $C = \int_a^b m(t)\mathrm{d}t$ and $M(z) = \int_a^z m(t)\mathrm{d}t$. Simple calculations show that

$$L^\star(1, z) = \frac{1}{(z-a)(b-z)}\left(M(z) - C\frac{z-a}{b-a}\right)^2.$$

Therefore, since $L^\star \equiv 0$ on $\mathcal{C}_A$ by assumption, we obtain

$$M(z) = C\,\frac{z-a}{b-a}\,.$$

This proves that $M(z)$ is linear in z, and that m is therefore constant on $[a, b]$.

Let us now examine the general multivariate case, where $A = \Pi_{j=1}^{p}[a_j, b_j] \subset [0,1]^p$. From the univariate analysis, we know that, for all $1 \leq j \leq p$, there exists a constant C_j such that

$$\int_{a_1}^{b_1} \dots \int_{a_p}^{b_p} m(\mathbf{x}) \mathrm{d}x_1 \dots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \dots \mathrm{d}x_p = C_j \,.$$

Since m is additive this implies that, for all j and all x_j,

$$m_j(x_j) = C_j - \int_{a_1}^{b_1} \dots \int_{a_p}^{b_p} \sum_{\ell \neq j} m_\ell(x_\ell) \mathrm{d}x_1 \dots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \dots \mathrm{d}x_p \,,$$

which does not depend upon x_i. This shows that m is constant on A. □

Proof of Lemma 1 Take $\xi > 0$ and fix $\mathbf{x} \in [0,1]^p$. Let θ be a realization of the random variable Θ. Since m is uniformly continuous, the result is clear if $\mathrm{diam}(A_k^\star(\mathbf{x}, \theta))$ tends to zero as k tends to infinity. Thus, in the sequel, it is assumed that $\mathrm{diam}(A_k^\star(\mathbf{x}, \theta))$ does not tend to zero. In that case, since $(A_k^\star(\mathbf{x}, \theta))_k$ is a decreasing sequence of compact sets, there exist $\mathbf{a}_\infty(\mathbf{x}, \theta) = (\mathbf{a}_\infty^{(1)}(\mathbf{x}, \theta), \dots, \mathbf{a}_\infty^{(p)}(\mathbf{x}, \theta)) \in [0,1]^p$ and $\mathbf{b}_\infty(\mathbf{x}, \theta) = (\mathbf{b}_\infty^{(1)}(\mathbf{x}, \theta), \dots, \mathbf{b}_\infty^{(p)}(\mathbf{x}, \theta)) \in [0,1]^p$ such that

$$\begin{aligned} \bigcap_{k=1}^{\infty} A_k^\star(\mathbf{x}, \theta) &= \prod_{j=1}^{p} [\mathbf{a}_\infty^{(j)}(\mathbf{x}, \theta), \mathbf{b}_\infty^{(j)}(\mathbf{x}, \theta)] \\ &\stackrel{\text{def}}{=} A_\infty^\star(\mathbf{x}, \theta). \end{aligned}$$

Since $\mathrm{diam}(A_k^\star(\mathbf{x}, \theta))$ does not tend to zero, there exists an index j' such that $\mathbf{a}_\infty^{(j')}(\mathbf{x}, \theta) < \mathbf{b}_\infty^{(j')}(\mathbf{x}, \theta)$ (i.e., the cell $A_\infty^\star(\mathbf{x}, \theta)$ is not reduced to one point). Let $A_k^\star(\mathbf{x}, \theta) \stackrel{\text{def}}{=} \prod_{j=1}^{p} [\mathbf{a}_k^{(j)}(\mathbf{x}, \theta), \mathbf{b}_k^{(j)}(\mathbf{x}, \theta)]$ be the cell containing $\mathbf{x}$ at level k. If the criterion $L^\star$ is identically zero for all cuts in $A_\infty^\star(\mathbf{x}, \theta)$ then m is constant on $A_\infty^\star(\mathbf{x}, \theta)$ according to Lemma 5.1. This implies that $\Delta(m, A_\infty^\star(\mathbf{x}, \theta)) = 0$. Thus, in that case, since m is uniformly continuous,

$$\lim_{k \to \infty} \Delta(m, A_k^\star(\mathbf{x}, \theta)) = \Delta(m, A_\infty^\star(\mathbf{x}, \theta)) = 0.$$

Let us now show by contradiction that $L^\star$ is almost surely necessarily null on the cuts of $A_\infty^\star(\mathbf{x}, \theta)$. In the rest of the proof, for all $k \in \mathbb{N}^\star$, we let $L_k^\star$ be the criterion $L^\star$ used in the cell $A_k^\star(\mathbf{x}, \theta)$, that is

$$\begin{aligned} L_k^\star(d) = \mathbb{V}[Y | \mathbf{X} \in A_k^\star(\mathbf{x}, \theta)] \\ - \mathbb{P}[\mathbf{X}^{(j)} < z \,|\, \mathbf{X} \in A_k^\star(\mathbf{x}, \theta)] \; \mathbb{V}[Y | \mathbf{X}^{(j)} < z, \mathbf{X} \in A_k^\star(\mathbf{x}, \theta)] \\ - \mathbb{P}[\mathbf{X}^{(j)} \geq z \,|\, \mathbf{X} \in A_k^\star(\mathbf{x}, \theta)] \; \mathbb{V}[Y | \mathbf{X}^{(j)} \geq z, \mathbf{X} \in A_k^\star(\mathbf{x}, \theta)], \end{aligned}$$

for all $d = (j, z) \in \mathcal{C}_{A_k^\star(\mathbf{x}, \theta)}$. If $L_\infty^\star$ is not identically zero, then there exists a cut $d_\infty(\mathbf{x}, \theta)$ in $\mathcal{C}_{A_\infty^\star(\mathbf{x}, \theta)}$ such that $L^\star(d_\infty(\mathbf{x}, \theta)) = c > 0$. Fix $\xi > 0$. By the uniform continuity of m, there exists $\delta_1 > 0$ such that

$$\sup_{\|\mathbf{w} - \mathbf{w}'\|_\infty \leq \delta_1} |m(\mathbf{w}) - m(\mathbf{w}')| \leq \xi \,.$$

Since $A_k^\star(\mathbf{x},\theta) \downarrow A_\infty^\star(\mathbf{x},\theta)$, there exists k_0 such that, for all $k \geq k_0$,

$$\max\left(\|\mathbf{a}_k(\mathbf{x},\theta) - \mathbf{a}_\infty(\mathbf{x},\theta)\|_\infty, \|\mathbf{b}_k(\mathbf{x},\theta) - \mathbf{b}_\infty(\mathbf{x},\theta)\|_\infty\right) \leq \delta_1\,. \tag{5.12}$$

Observe that, for all $k \in \mathbb{N}^\star$, $\mathbb{V}[Y|\mathbf{X} \in A_{k+1}^\star(\mathbf{x},\theta)] < \mathbb{V}[Y|\mathbf{X} \in A_k^\star(\mathbf{x},\theta)]$. Thus,

$$\underline{L}_k^\star := \sup_{\substack{d \in \mathcal{C}_{A_k(\mathbf{x},\theta)} \\ d^{(1)} \in \mathcal{M}_{\text{try}}}} L_k^\star(d) \leq \xi\,. \tag{5.13}$$

From inequality (5.12), we deduce that

$$\left|\mathbb{E}[m(\mathbf{X})|\mathbf{X} \in A_k^\star(\mathbf{x},\theta)] - \mathbb{E}[m(\mathbf{X})|\mathbf{X} \in A_\infty^\star(\mathbf{x},\theta)]\right| \leq \xi.$$

Consequently, there exists a constant $C > 0$ such that, for all $k \geq k_0$ and all cuts $d \in \mathcal{C}_{A_\infty^\star(\mathbf{x},\theta)}$,

$$|L_k^\star(d) - L_\infty^\star(d)| \leq C\xi^2. \tag{5.14}$$

Let $k_1 \geq k_0$ be the first level after k_0 at which the direction $d_\infty^{(1)}(\mathbf{x},\theta)$ is amongst the m_{try} selected coordinates. Almost surely, $k_1 < \infty$. Thus, by the definition of $d_\infty(\mathbf{x},\theta)$ and inequality (5.14),

$$c - C\xi^2 \leq L_\infty^\star(d_\infty(\mathbf{x},\theta)) - C\xi^2 \leq L_k^\star(d_\infty(\mathbf{x},\theta)),$$

which implies that $c - C\xi^2 \leq \underline{L}_k^\star$. Hence, using inequality (5.13), we have

$$c - C\xi^2 \leq \underline{L}_k^\star \leq \xi,$$

which is absurd, since $c > 0$ is fixed and ξ is arbitrarily small. Thus, by Lemma 5.1, m is constant on $A_\infty^\star(\mathbf{x},\theta)$. This implies that $\Delta(m, A_k^\star(\mathbf{x},\Theta)) \to 0$ as $k \to \infty$.

5.6.2 Proof of Lemma 2

We start by proving Lemma 2 in the case $k = 1$, i.e., when the first cut is performed at the root of a tree. Since in that case $L_{n,1}(\mathbf{x},\cdot)$ does not depend on $\mathbf{x}$, we simply write $L_{n,1}(\cdot)$ instead of $L_{n,1}(\mathbf{x},\cdot)$.

Proof of Lemma 2 in the case $k = 1$. Fix $\alpha, \rho > 0$. Observe that if two cuts d_1, d_2 satisfy $\|d_1 - d_2\|_\infty < 1$, then the cut directions are the same, i.e., $d_1^{(1)} = d_2^{(1)}$. Using this fact and symmetry arguments, we just need to prove Lemma 2 when the cuts are performed along the first dimension. In other words, we only need to prove that

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{|x_1 - x_2| \leq \delta} |L_{n,1}(1,x_1) - L_{n,1}(1,x_2)| > \alpha\right] \leq \rho/p. \tag{5.15}$$

Preliminary results Letting $Z_i = \max_{1\leq i\leq n} |\varepsilon_i|$, simple calculations show that

$$\mathbb{P}\left[Z_i \geq t\right] = 1 - \exp\Big(n \ln\left(1 - 2\mathbb{P}\left[\varepsilon_1 \geq t\right]\right)\Big).$$

The last probability can be upper bounded by using the following standard inequality on Gaussian tail:

$$\mathbb{P}\left[\varepsilon_1 \geq t\right] \leq \frac{\sigma}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Consequently, there exists a constant $C_\rho > 0$ and $N_1 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $n > N_1$,

$$\max_{1\leq i\leq n} |\varepsilon_i| \leq C_\rho \sqrt{\log n}\,. \tag{5.16}$$

Besides, by simple calculations on Gaussian tail, for all $n \in \mathbb{N}^\star$, we have

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\right| \geq \alpha\right] \leq \frac{\sigma}{\alpha\sqrt{n}} \exp\left(-\frac{\alpha^2 n}{2\sigma^2}\right).$$

Since there are, at most, n^2 sets of the form $\{i : X_i \in [a_n, b_n]\}$ for $0 \leq a_n < b_n \leq 1$, we deduce from the last inequality and the union bound, that there exists $N_2 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $n > N_2$ and all $0 \leq a_n < b_n \leq 1$ satisfying $N_n([a_n, b_n] \times [0,1]^{p-1}) > \sqrt{n}$,

$$\Big|\frac{1}{N_n([a_n, b_n] \times [0,1]^{p-1})} \sum_{\substack{i: X_i \in [a_n, b_n] \\ \times [0,1]^{p-1}}} \varepsilon_i\Big| \leq \alpha. \tag{5.17}$$

By the Glivenko-Cantelli theorem, there exists $N_3 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $0 \leq a < b \leq 1$, and all $n > N_3$,

$$(b - a - \delta^2)n \leq N_n([a,b] \times [0,1]^{p-1}) \leq (b - a + \delta^2)n. \tag{5.18}$$

Throughout the proof, we assume to be on the event where assertions (5.16)-(5.18) hold, which occurs with probability $1-3\rho$, for all $n > N$, where $N = \max(N_1, N_2, N_3)$.

Take $x_1, x_2 \in [0,1]$ such that $|x_1 - x_2| \leq \delta$ and assume, without loss of generality, that $x_1 < x_2$. In the remainder of the proof, we will need the following quantities (see Figure 5.1 for an illustration in dimension two):

$$\begin{cases} A_{L,\sqrt{\delta}} = [0, \sqrt{\delta}] \times [0,1]^{p-1} \\ A_{R,\sqrt{\delta}} = [1-\sqrt{\delta}, 1] \times [0,1]^{p-1} \\ A_{C,\sqrt{\delta}} = [\sqrt{\delta}, 1-\sqrt{\delta}] \times [0,1]^{p-1}. \end{cases}$$

Similarly, we define

$$\begin{cases} A_{L,1} &= [0, x_1] \times [0,1]^{p-1} \\ A_{R,1} &= [x_1, 1] \times [0,1]^{p-1} \\ A_{L,2} &= [0, x_2] \times [0,1]^{p-1} \\ A_{R,2} &= [x_2, 1] \times [0,1]^{p-1} \\ A_C &= [x_1, x_2] \times [0,1]^{p-1}. \end{cases}$$

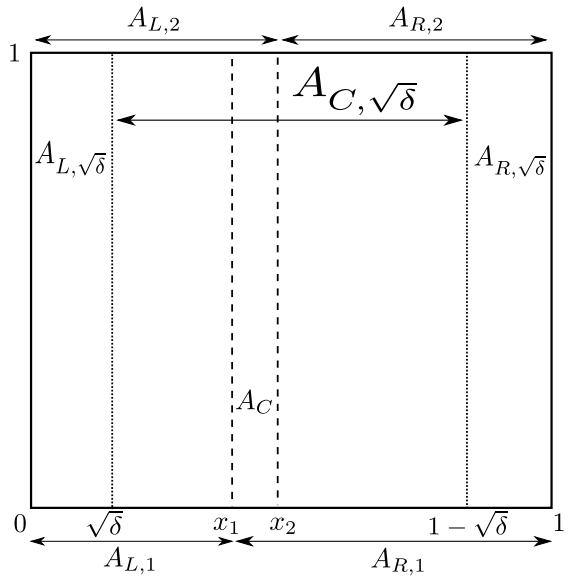

Figure 5.1: Illustration of the notation in dimension $p = 2$.

Recall that, for any cell A, $\bar{Y}_A$ is the mean of the Y_i's falling in A and $N_n(A)$ is the number of data points in A. To prove (5.15), five cases are to be considered, depending upon the positions of x_1 and x_2. We repeatedly use the decomposition

$$L_{n,1}(1, x_1) - L_{n,1}(1, x_2) = J_1 + J_2 + J_3,$$

where

$$J_1 = \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} (Y_i - \bar{Y}_{A_{L,1}})^2 - \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} (Y_i - \bar{Y}_{A_{L,2}})^2,$$

$$J_2 = \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} \in [x_1, x_2]} (Y_i - \bar{Y}_{A_{R,1}})^2 - \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} \in [x_1, x_2]} (Y_i - \bar{Y}_{A_{L,2}})^2,$$

$$\text{and} \quad J_3 = \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} \geq x_2} (Y_i - \bar{Y}_{A_{R,1}})^2 - \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} \geq x_2} (Y_i - \bar{Y}_{A_{R,2}})^2.$$

First case Assume that $x_1, x_2 \in A_{C,\sqrt{\delta}}$. Since $N_n(A_{L,2}) > N_n(A_{L,\sqrt{\delta}}) > \sqrt{n}$ for all $n > N$, we have, according to inequalities (5.17),

$$|\bar{Y}_{A_{L,2}}| \leq \|m\|_\infty + \alpha \quad \text{and} \quad |\bar{Y}_{A_{R,1}}| \leq \|m\|_\infty + \alpha.$$

Therefore

$$
\begin{aligned}
|J_2| &= 2\,|\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{R,1}}| \times \frac{1}{n}\Big|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\left(Y_i - \frac{\bar{Y}_{A_{L,2}} + \bar{Y}_{A_{R,1}}}{2}\right)\Big| \\
&\le 4(\|m\|_\infty + \alpha)\Bigg(\frac{(\|m\|_\infty + \alpha)N_n(A_C)}{n} + \frac{1}{n}\Big|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} m(\mathbf{X}_i)\Big| \\
&\qquad\qquad + \frac{1}{n}\Big|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\Big|\Bigg) \\
&\le 4(\|m\|_\infty + \alpha)\Bigg((\delta+\delta^2)(\|m\|_\infty + \alpha) + \|m\|_\infty(\delta+\delta^2) \\
&\qquad\qquad + \frac{1}{n}\Big|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\Big|\Bigg).
\end{aligned}
$$

If $N_n(A_C) \ge \sqrt{n}$, we obtain

$$
\frac{1}{n}\Big|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\Big| \le \frac{1}{N_n(A_C)}\Big|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\Big| \le \alpha \quad \text{(according to (5.17))}
$$

or, if $N_n(A_C) < \sqrt{n}$, we have

$$
\frac{1}{n}\Big|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\Big| \le \frac{C_\rho\sqrt{\log n}}{\sqrt{n}} \quad \text{(according to (5.16))}.
$$

Thus, for all n large enough,

$$
|J_2| \le 4(\|m\|_\infty + \alpha)\Big((\delta+\delta^2)(2\|m\|_\infty + \alpha) + \alpha\Big). \tag{5.19}
$$

With respect to J_1, observe that

$$\begin{aligned}
|\bar{Y}_{A_{L,1}} - \bar{Y}_{A_{L,2}}| &= \left| \frac{1}{N_n(A_{L,1})} \sum_{i:\mathbf{X}_i^{(1)} < x_1} Y_i - \frac{1}{N_n(A_{L,2})} \sum_{i:\mathbf{X}_i^{(1)} < x_2} Y_i \right| \\
&\leq \left| \frac{1}{N_n(A_{L,1})} \sum_{i:\mathbf{X}_i^{(1)} < x_1} Y_i - \frac{1}{N_n(A_{L,2})} \sum_{i:\mathbf{X}_i^{(1)} < x_1} Y_i \right| \\
&\quad + \left| \frac{1}{N_n(A_{L,2})} \sum_{i:\mathbf{X}_i^{(1)} \in [x_1, x_2]} Y_i \right| \\
&\leq \left| 1 - \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \right| \times \frac{1}{N_n(A_{L,1})} \times \left| \sum_{i:\mathbf{X}_i^{(1)} < x_1} Y_i \right| \\
&\quad + \frac{1}{N_n(A_{L,2})} \left| \sum_{i:\mathbf{X}_i^{(1)} \in [x_1, x_2]} Y_i \right|.
\end{aligned}$$

Since $N_n(A_{L,2}) - N_n(A_{L,1}) \leq n(\delta + \delta^2)$, we obtain

$$1 - \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \leq \frac{n(\delta + \delta^2)}{N_n(A_{L,2})} \leq \frac{\delta + \delta^2}{\sqrt{\delta} - \delta^2} \leq 4\sqrt{\delta},$$

for all δ small enough, which implies that

$$\begin{aligned}
|\bar{Y}_{A_{L,1}} - \bar{Y}_{A_{L,2}}| &\leq \frac{4\sqrt{\delta}}{N_n(A_{L,1})} \left| \sum_{i:\mathbf{X}_i^{(1)} < x_1} Y_i \right| \\
&\quad + \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \times \frac{1}{N_n(A_{L,1})} \left| \sum_{i:\mathbf{X}_i^{(1)} \in [x_1, x_2]} Y_i \right| \\
&\leq 4\sqrt{\delta}(\|m\|_\infty + \alpha) + \frac{N_n(A_{L,1})}{N_n(A_{L,2})}(\|m\|_\infty \delta + \alpha) \\
&\leq 5(\|m\|_\infty \sqrt{\delta} + \alpha).
\end{aligned}$$

Thus,

$$\begin{aligned}
|J_1| &= \left| \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} (Y_i - \bar{Y}_{A_{L,1}})^2 - \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} (Y_i - \bar{Y}_{A_{L,2}})^2 \right| \\
&= \left| (\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{L,1}}) \times \frac{2}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} \left(Y_i - \frac{\bar{Y}_{A_{L,1}} + \bar{Y}_{A_{L,2}}}{2} \right) \right| \\
&\leq |\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{L,1}}|^2 \\
&\leq 25(\|m\|_\infty \sqrt{\delta} + \alpha)^2. \qquad (5.20)
\end{aligned}$$

The term J_3 can be bounded with the same arguments.

Finally, by (5.19) and (5.20), for all $n > N$, and all δ small enough, we conclude that

$$\begin{aligned}|L_n(1,x_1) - L_n(1,x_2)| &\leq 4(\|m\|_\infty + \alpha)\Big((\delta+\delta^2)(2\|m\|_\infty+\alpha)+\alpha\Big) \\ &\quad + 25(\|m\|_\infty\sqrt{\delta}+\alpha)^2 \\ &\leq \alpha.\end{aligned}$$

Second case Assume that $x_1, x_2 \in A_{L,\sqrt{\delta}}$. With the same arguments as above, one proves that

$$\begin{aligned}|J_1| &\leq \max\left(4(\sqrt{\delta}+\delta^2)(\|m\|_\infty+\alpha)^2, \alpha\right), \\ |J_2| &\leq \max(4(\|m\|_\infty+\alpha)(2\delta\|m\|_\infty+2\alpha), \alpha), \\ |J_3| &\leq 25(\|m\|_\infty\sqrt{\delta}+\alpha)^2.\end{aligned}$$

Consequently, for all n large enough,

$$|L_n(1,x_1) - L_n(1,x_2)| = J_1 + J_2 + J_3 \leq 3\alpha.$$

The other cases $\{x_1, x_2 \in A_{R,\sqrt{\delta}}\}$, $\{x_1, x_2 \in A_{L,\sqrt{\delta}} \times A_{C,\sqrt{\delta}}\}$, and $\{x_1, x_2 \in A_{C,\sqrt{\delta}} \times A_{R,\sqrt{\delta}}\}$ can be treated in the same way. Details are omitted. □

Proof of Lemma 2. We proceed similarly as in the proof of the case $k = 1$. Here, we establish the result for $k = 2$ and $p = 2$ only. Extensions are easy and left to the reader.

Preliminary results Fix $\rho > 0$. At first, it should be noted that there exists $N_1 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $n > N_0$ and all $A_n = [a_n^{(1)}, b_n^{(1)}] \times [a_n^{(2)}, b_n^{(2)}] \subset [0,1]^2$ satisfying $N_n(A_n) > \sqrt{n}$, we have

$$\Big|\frac{1}{N_n(A_n)} \sum_{i:X_i \in A_n} \varepsilon_i\Big| \leq \alpha, \tag{5.21}$$

and

$$\frac{1}{N_n(A_n)} \sum_{i:X_i \in A_n} \varepsilon_i^2 \leq \tilde{\sigma}^2, \tag{5.22}$$

where $\tilde{\sigma}^2$ is a positive constant, depending only on ρ. Inequality (5.22) is a straightforward consequence of the following inequality [see, e.g., Laurent and Massart, 2000], which is valid for all $n \in \mathbb{N}^\star$:

$$\mathbb{P}\left[\chi^2(n) \geq 5n\right] \leq \exp(-n).$$

Throughout the proof, we assume to be on the event where assertions (5.16), (5.18), (5.21)-(5.22) hold, which occurs with probability $1 - 3\rho$, for all n large enough. We also assume that $d_1 = (1, x_1)$ and $d_2 = (2, x_2)$ (see Figure 5.2). The other cases can be treated similarly.

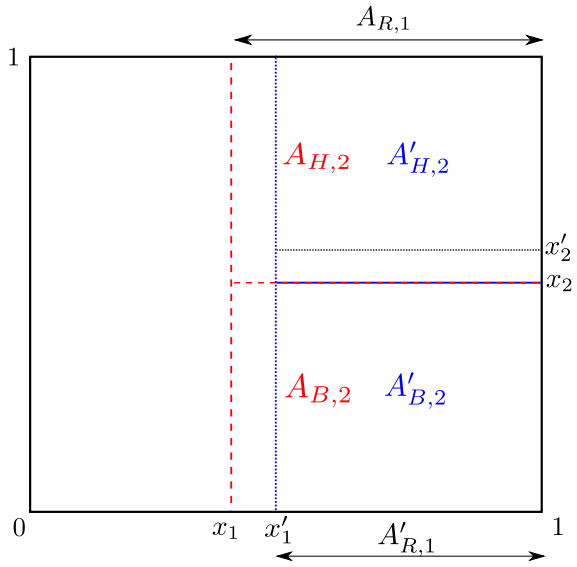

Figure 5.2: An example of cells in dimension $p = 2$.

Main argument Let $d'_1 = (1, x'_1)$ and $d'_2 = (2, x'_2)$ be such that $|x_1 - x'_1| < \delta$ and $|x_2 - x'_2| < \delta$. Then the CART-split criterion $L_{n,2}$ writes

$$\begin{aligned} L_n(d_1, d_2) = \frac{1}{N_n(A_{R,1})} \sum_i (Y_i - \bar{Y}_{A_{R,1}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ - \frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ - \frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} \leq x_2} (Y_i - \bar{Y}_{A_{B,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1}. \end{aligned}$$

Clearly,

$$L_n(d_1, d_2) - L_n(d'_1, d'_2) = L_n(d_1, d_2) - L_n(d'_1, d_2) + L_n(d'_1, d_2) - L_n(d'_1, d'_2).$$

We have (Figure 5.2):

$$\begin{aligned}
L_n(d_1,d_2) - L_n(d_1',d_2) = &\left[\frac{1}{N_n(A_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}>x_2}(Y_i-\bar{Y}_{A_{H,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1}\right.\\
&\left.-\frac{1}{N_n(A'_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}>x_2}(Y_i-\bar{Y}_{A'_{H,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1'}\right]\\
&+\left[\frac{1}{N_n(A_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}\leq x_2}(Y_i-\bar{Y}_{A_{B,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1}\right.\\
&\left.-\frac{1}{N_n(A'_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}\leq x_2}(Y_i-\bar{Y}_{A'_{B,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1'}\right]\\
&\stackrel{\text{def}}{=} A_1 + B_1.
\end{aligned}$$

The term A_1 can be rewritten as $A_1 = A_{1,1} + A_{1,2} + A_{1,3}$, where

$$\begin{aligned}
A_{1,1} = &\frac{1}{N_n(A_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}>x_2}(Y_i-\bar{Y}_{A_{H,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1'}\\
&-\frac{1}{N_n(A_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}>x_2}(Y_i-\bar{Y}_{A'_{H,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1'},\\
A_{1,2} = &\frac{1}{N_n(A_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}>x_2}(Y_i-\bar{Y}_{A'_{H,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1'}\\
&-\frac{1}{N_n(A'_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}>x_2}(Y_i-\bar{Y}_{A'_{H,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}>x_1'},\\
\text{and}\quad A_{1,3} = &\frac{1}{N_n(A_{R,1})}\sum_{i:\mathbf{X}_i^{(2)}>x_2}(Y_i-\bar{Y}_{A_{H,2}})^2\mathbb{1}_{\mathbf{X}_i^{(1)}\in[x_1,x_1']}.
\end{aligned}$$

Easy calculations show that

$$A_{1,1} = \frac{N_n(A'_{H,2})}{N_n(A_{R,1})}(\bar{Y}_{A'_{H,2}} - \bar{Y}_{A_{H,2}})^2,$$

which implies, with the same arguments as in the proof for $k=1$, that $A_{1,1} \to 0$ as $n \to \infty$. With respect to $A_{1,2}$ and $A_{1,3}$, we write

$$\max(A_{1,2}, A_{1,3}) \leq \max(C_\rho \frac{\log n}{\sqrt{n}}, 2(\tilde{\sigma}^2 + 4\|m\|_\infty^2 + \alpha^2)\frac{\sqrt{\delta}}{\xi}).$$

Thus, $A_{1,2} \to 0$ and $A_{1,3} \to 0$ as $n \to \infty$. Collecting bounds, we conclude that $A_1 \to 0$. One proves with similar arguments that $B_1 \to 0$ and, consequently, that $L_n(d_1', d_2) - L_n(d_1', d_2') \to 0$. □

5.6.3 Proof of Lemma 3

We prove by induction that, for all k, with probability $1-\rho$, for all $\xi > 0$ and all n large enough,

$$d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \leq \xi.$$

Call this property H_k. Fix $k > 1$ and assume that H_{k-1} is true. For all $\mathbf{d}_{k-1} \in \mathcal{A}_{k-1}(\mathbf{X})$, let

$$\hat{d}_{k,n}(\mathbf{d}_{k-1}) \in \arg\min_{d_k} L_n(\mathbf{X}, \mathbf{d}_{k-1}, d_k),$$

and

$$d_k^\star(\mathbf{d}_{k-1}) \in \arg\min_{d_k} L^\star(\mathbf{X}, \mathbf{d}_{k-1}, d_k),$$

where the minimum is evaluated, as usual, over $\{d_k \in \mathcal{C}_{A(\mathbf{X}, \mathbf{d}_{k-1})} : d_k^{(1)} \in \mathcal{M}_{\text{try}}\}$. Fix $\rho > 0$. In the rest of the proof, we assume Θ to be fixed and we omit the dependence on Θ.

Preliminary result We momentarily consider $\mathbf{x} \in [0,1]^p$. Note that, for all $\mathbf{d}_{k-1}$,

$$\begin{aligned}
&L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) \\
&\quad \leq L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) \\
&\qquad \text{(by definition of } d_k^\star(\mathbf{d}_{k-1})) \\
&\quad \leq L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) \\
&\qquad \text{(by definition of } \hat{d}_{k,n}(\mathbf{d}_{k-1})).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left|L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\right| \\
&\quad \leq \max\Big(\left|L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1}))\right|, \\
&\qquad\qquad \left|L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\right|\Big) \\
&\quad \leq \sup_{d_k} \left|L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k)\right|.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&|L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))| \\
&\quad \leq |L^\star(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1}))| \\
&\qquad + |L_n(\mathbf{x}, \mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))| \\
&\quad \leq 2 \sup_{d_k} |L_n(\mathbf{x}, \mathbf{d}_{k-1}, d_k) - L^\star(\mathbf{x}, \mathbf{d}_{k-1}, d_k)| \\
&\quad = 2 \sup_{d_k} |L_n(\mathbf{x}, \mathbf{d}_k) - L^\star(\mathbf{x}, \mathbf{d}_k)|. \qquad (5.23)
\end{aligned}$$

Let $\bar{\mathcal{A}}_k^\xi(\mathbf{x}) = \{\mathbf{d}_k : \mathbf{d}_{k-1} \in \mathcal{A}_{k-1}^\xi(\mathbf{x})\}$. So, taking the supremum on both sides of (5.23) leads to

$$\begin{aligned}\sup_{\mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^\xi(\mathbf{x})} & |L^\star(\mathbf{x},\mathbf{d}_{k-1},\hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x},\mathbf{d}_{k-1},d_k^\star(\mathbf{d}_{k-1}))| \\ &\leq 2 \sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^\xi(\mathbf{x})} |L_n(\mathbf{x},\mathbf{d}_k) - L^\star(\mathbf{x},\mathbf{d}_k)|. \end{aligned} \tag{5.24}$$

By Lemma 3, for all $\xi' > 0$, one can find $\delta > 0$ such that, for all n large enough,

$$\mathbb{P}\left[\sup_{\mathbf{x}\in[0,1]^p} \sup_{\substack{\|\mathbf{d}_k-\mathbf{d}_k'\|_\infty\leq\delta \\ \mathbf{d}_k,\mathbf{d}_k'\in\bar{\mathcal{A}}_k^\xi(\mathbf{x})}} |L_n(\mathbf{x},\mathbf{d}_k) - L_n(\mathbf{x},\mathbf{d}_k')| \leq \xi'\right] \geq 1-\rho. \tag{5.25}$$

Now, let $\mathcal{G}$ be a regular grid of $[0,1]^p$ whose grid step equal to $\xi/2$. Note that, for all $\mathbf{x}\in\mathcal{G}$, $\bar{\mathcal{A}}_k^\xi(\mathbf{x})$ is compact. Thus, for all $\mathbf{x}\in\mathcal{G}$, there exists a finite subset $\mathcal{C}_{\delta,\mathbf{x}} = \{c_{j,\mathbf{x}} : 1\leq j\leq p\}$ such that, for all $\mathbf{d}_k\in\bar{\mathcal{A}}_k^\xi(\mathbf{x})$, $d_\infty(\mathbf{d}_k,\mathcal{C}_{\delta,\mathbf{x}})\leq\delta$. Set $\xi'>0$. Observe that, since the subset $\cup_{\mathbf{x}\in\mathcal{G}}\mathcal{C}_{\delta,\mathbf{x}}$ is finite, one has, for all n large enough,

$$\sup_{\mathbf{x}\in\mathcal{G}} \sup_{c_{j,\mathbf{x}}\in\mathcal{C}_{\delta,\mathbf{x}}} |L_n(\mathbf{x},c_{j,\mathbf{x}}) - L^\star(\mathbf{x},c_{j,\mathbf{x}})| \leq \xi'. \tag{5.26}$$

Hence, for all n large enough,

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{G}} \sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^\xi(\mathbf{x})} |L_n(\mathbf{x},\mathbf{d}_k) - L^\star(\mathbf{x},\mathbf{d}_k)| \leq \sup_{\mathbf{x}\in\mathcal{G}} \sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^\xi(\mathbf{x})} \Big(& |L_n(\mathbf{x},\mathbf{d}_k) - L_n(\mathbf{x},c_{j,\mathbf{x}})| \\ & + |L_n(\mathbf{x},c_{j,\mathbf{x}}) - L^\star(\mathbf{x},c_{j,\mathbf{x}})| + |L^\star(\mathbf{x},c_{j,\mathbf{x}}) - L^\star(\mathbf{x},\mathbf{d}_k)|\Big),\end{aligned}$$

where $c_{j,\mathbf{x}}$ satisfies $\|c_{j,\mathbf{x}}-\mathbf{d}_k\|_\infty\leq\delta$. Using inequalities (5.25) and (5.26), with probability $1-\rho$, we obtain, for all n large enough,

$$\sup_{\mathbf{x}\in\mathcal{G}} \sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^\xi(\mathbf{x})} |L_n(\mathbf{x},\mathbf{d}_k) - L^\star(\mathbf{x},\mathbf{d}_k)| \leq 3\xi'.$$

Finally, by inequality (5.24), with probability $1-\rho$, for all n large enough,

$$\sup_{\mathbf{x}\in\mathcal{G}} \sup_{\mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^\xi(\mathbf{x})} |L^\star(\mathbf{x},\mathbf{d}_{k-1},\hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{x},\mathbf{d}_{k-1},d_k^\star(\mathbf{d}_{k-1}))| \leq 6\xi'. \tag{5.27}$$

Hereafter, to simplify, we assume that, for any given $(k-1)$-tuple of theoretical cuts, there is only one theoretical cut at level k, and leave the general case as an easy adaptation. Thus, we can define unambiguously

$$d_k^\star(\mathbf{d}_{k-1}) = \arg\min_{d_k} L^\star(\mathbf{d}_{k-1},d_k).$$

Fix $\xi'' > 0$. From inequality (5.27), by evoking the equicontinuity of L_n and the compactness of $\mathcal{U} = \{(\mathbf{x}, \mathbf{d}_{k-1}) : \mathbf{x} \in \mathcal{G}, \mathbf{d}_{k-1} \in \mathcal{A}_{k-1}^{\xi}(\mathbf{x})\}$, we deduce that, with probability $1-\rho$, for all n large enough,

$$\sup_{(\mathbf{x},\mathbf{d}_{k-1})\in\mathcal{U}} d_\infty\Big(\hat{d}_{k,n}(\mathbf{d}_{k-1}), d_k^\star(\mathbf{d}_{k-1})\Big) \leq \xi''. \tag{5.28}$$

Besides,

$$\mathbb{P}\big[(\mathbf{X}, \hat{\mathbf{d}}_{k-1,n}(\mathbf{X})) \in \mathcal{U}\big] = \mathbb{E}\big[\mathbb{P}\big[(\mathbf{X}, \hat{\mathbf{d}}_{k-1,n}(\mathbf{X})) \in \mathcal{U}|\mathcal{D}_n\big]\big] \geq 1 - 2^{k-1}\xi. \tag{5.29}$$

In the rest of the proof, we consider $\xi \leq \rho/2^{k-1}$, which, by inequalities (5.28) and (5.29), leads to

$$\mathbb{P}\big[\sup_{(\mathbf{x},\mathbf{d}_{k-1})\in\mathcal{U}} d_\infty\Big(\hat{d}_{k,n}(\mathbf{d}_{k-1}), d_k^\star(\mathbf{d}_{k-1})\Big) \leq \xi'', (\mathbf{X}, \hat{\mathbf{d}}_{k-1,n}(\mathbf{X})) \in \mathcal{U}\big] \geq 1 - 2\rho.$$

This implies, with probability $1-2\rho$, for all n large enough,

$$d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\hat{\mathbf{d}}_{k-1,n})\Big) \leq \xi''. \tag{5.30}$$

Main argument Now, using triangle inequality,

$$\begin{aligned} d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^\star\Big) \leq\ & d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\hat{\mathbf{d}}_{k-1,n})\Big) \\ & + d_\infty\Big(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^\star\Big). \end{aligned} \tag{5.31}$$

Thus, we just have to show that $d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^\star) \to 0$ in probability as $n \to \infty$, and the proof will be complete. To avoid confusion, we let $\{\mathbf{d}_{k-1}^{\star,i} : i \in \mathcal{I}\}$ be the set of best first $(k-1)$-th theoretical cuts (which can be either countable or not). With this notation, $d_k^\star(\mathbf{d}_{k-1}^{\star,i})$ is the k-th theoretical cuts given that the $(k-1)$ previous ones are $\mathbf{d}_{k-1}^{\star,i}$. For simplicity, let

$$L^{i,\star}(\mathbf{x}, d_k) = L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, d_k) \quad \text{and} \quad \hat{L}^\star(\mathbf{x}, d_k) = L_k^\star(\mathbf{x}, \hat{\mathbf{d}}_{k-1,n}, d_k).$$

As before,

$$d_k^\star(\mathbf{d}_{k-1}^{\star,i}) \in \arg\min_{d_k} L^{i,\star}(\mathbf{x}, d_k) \quad \text{and} \quad d_k^\star(\hat{\mathbf{d}}_{k-1,n}) \in \arg\min_{d_k} \hat{L}^\star(\mathbf{x}, d_k).$$

Clearly, the result will be proved if we establish that,

$$\inf_{i\in\mathcal{I}} d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \to 0, \quad \text{in probability,} \quad \text{as } n \to \infty.$$

Note that, for all $\mathbf{x} \in \mathcal{G}$, $\bar{\mathcal{A}}_k^{\xi}(\mathbf{x})$ is compact. Thus, for all $\mathbf{x} \in \mathcal{G}$, there exists a finite subset $\mathcal{C}'_{\delta,\mathbf{x}} = \{c'_{j,\mathbf{x}} : 1 \leq j \leq p\}$ such that, for all d_k, $d_\infty(d_k, \mathcal{C}'_{\delta,\mathbf{x}}) \leq \delta$. Hence, with probability $1-\rho$,

for all n large enough,

$$\begin{aligned}|\hat{L}^\star(\mathbf{x}, d_k) - L^{i,\star}(\mathbf{x}, d_k)| &\leq |\hat{L}^\star(\mathbf{x}, d_k) - \hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}})| \\ &\quad + |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})| \\ &\quad + |L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, d_k)| \\ &\leq 2\xi' + |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})| \\ &\quad \text{(by the continuity of } L_k^\star).\end{aligned}$$

Therefore, as in inequality (5.24), with probability $1-\rho$, for all i and all n large enough,

$$\begin{aligned}|L^{i,\star}(\mathbf{x}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{x}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| &\leq 2 \sup_{d_k} |\hat{L}^\star(\mathbf{x}, d_k) - L^{i,\star}(\mathbf{x}, d_k)| \\ &\leq 4\xi' + 2 \max_j |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})|.\end{aligned}$$

Taking the infimum over all i, we obtain

$$\begin{aligned}\inf_i |L^{i,\star}(\mathbf{x}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{x}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| &\leq 4\xi' \\ &\quad + 2 \inf_i \max_j |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})|. \qquad (5.32)\end{aligned}$$

Introduce ω, the modulus of continuity of $L_k^\star$:

$$\omega(\mathbf{x}, \delta) = \sup_{\|\mathbf{d}-\mathbf{d}'\|_\infty \leq \delta} |L_k^\star(\mathbf{x}, \mathbf{d}) - L_k^\star(\mathbf{x}, \mathbf{d}')|.$$

Observe that, since $L_k^\star(\mathbf{x}, \cdot)$ is uniformly continuous, $\omega(\delta) \to 0$ as $\delta \to 0$. Hence, for all n large enough,

$$\begin{aligned}&\inf_i \max_j |\hat{L}^\star(\mathbf{x}, c'_{j,\mathbf{x}}) - L^{i,\star}(\mathbf{x}, c'_{j,\mathbf{x}})| \\ &= \inf_i \max_j |L_k^\star(\mathbf{x}, \hat{\mathbf{d}}_{k-1,n}, c'_{j,\mathbf{x}}) - L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, c'_{j,\mathbf{x}})| \\ &\leq \inf_i \omega(\mathbf{x}, \|\hat{\mathbf{d}}_{k-1,n} - \mathbf{d}_{k-1}^{\star,i}\|_\infty) \\ &\leq \xi', \qquad (5.33)\end{aligned}$$

since, by assumption H_{k-1}, $\inf_i \|\hat{\mathbf{d}}_{k-1,n} - \mathbf{d}_{k-1}^{\star,i}\|_\infty \to 0$. Therefore, combining (5.32) and (5.33), with probability $1-\rho$, for all n large enough,

$$\inf_i |L^{i,\star}(\mathbf{X}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{X}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \leq 6\xi.$$

Finally, by Technical Lemma 5.2 below, with probability $1-\rho$, for all n large enough,

$$\inf_i d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \leq \xi''. \qquad (5.34)$$

Plugging inequality (5.34) and (5.30) into (5.31), we conclude that, with probability $1-3\rho$, for all n large enough,

$$d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^\star\Big) \leq 2\xi'',$$

which proves H_k. Property H_1 can be proved in the same way.

Technical Lemma 5.2. *For all $\delta, \rho > 0$, there exists $\xi > 0$ such that, if, with probability $1-\rho$,*

$$\inf_i |L^{i,\star}(\mathbf{X}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{X}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \leq \xi,$$

then, with probability $1-\rho$,

$$\inf_i d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \leq \delta. \tag{5.35}$$

Proof of Technical Lemma 5.2. Fix $\rho > 0$. Note that, for all $\delta > 0$, there exists $\xi > 0$ such that,

$$\inf_{\mathbf{x}\in[0,1]^p} \inf_i \inf_{y: d_\infty(y, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \geq \delta} |L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) - L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, y)| \geq \xi.$$

To see this, assume that one can find $\delta > 0$ such that, for all $\xi > 0$, there exist $i_\xi, y_\xi, \mathbf{x}_\xi$ satisfying

$$|L_k^\star(\mathbf{x}_\xi, \mathbf{d}_{k-1}^{\star,i_\xi}, d_k^\star(\mathbf{d}_{k-1}^{\star,i_\xi})) - L_k^\star(\mathbf{x}_\xi, \mathbf{d}_{k-1}^{\star,i_\xi}, y_\xi)| \leq \xi,$$

with $d_\infty(y_\xi, d_k^\star(\mathbf{d}_{k-1}^{\star,i_\xi})) \geq \delta$. Recall that $\{\mathbf{d}_{k-1}^{\star,i} : i \in \mathbb{N}\}$, $\{d_k^\star(\mathbf{d}_{k-1}^{\star,i}) : i \in \mathbb{N}\}$ are compact. Then, letting $\xi_p = 1/p$, we can extract three sequences $\mathbf{d}_{k-1}^{\star,i_p} \to \mathbf{d}_{k-1}$, $d_k^\star(\mathbf{d}_{k-1}^{\star,i_p}) \to d_k$ and $y_{\xi_{i_p}} \to y$ as $p \to \infty$ such that

$$L_k^\star(\mathbf{d}_{k-1}, d_k) = L_k^\star(\mathbf{d}_{k-1}, y), \tag{5.36}$$

and $d_\infty(y, d_k) \geq \delta$. Since we assume that given the $(k-1)$-th first cuts $\mathbf{d}_{k-1}$, there is only one best cut d_k, equation (5.36) implies that $y = d_k$, which is absurd.

Now, to conclude the proof, fix $\delta > 0$ and assume that, with probability $1-\rho$,

$$\inf_i d_\infty(d_k^\star(\mathbf{d}_{k-1}^{\star,i}), d_k^\star(\hat{\mathbf{d}}_{k-1,n})) \geq \delta.$$

Thus, with probability $1-\rho$,

$$\begin{aligned}
&\inf_i |L^{i,\star}(\mathbf{X}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(\mathbf{X}, d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \\
&\quad = \inf_i |L_k^\star(\mathbf{X}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L_k^\star(\mathbf{X}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star)| \\
&\quad \geq \inf_{\mathbf{x}\in[0,1]^p} \inf_i \inf_{d_\infty(y, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \geq \delta} |L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, y) - L_k^\star(\mathbf{x}, \mathbf{d}_{k-1}^{\star,i}, d_k^\star)| \\
&\quad \geq \xi,
\end{aligned}$$

which, by contraposition, concludes the proof. □

Proof of Proposition 1. Fix $k \in \mathbb{N}^\star$ and $\rho, \xi > 0$. According to Lemma 3, with probability $1-\rho$, for all n large enough, there exists a sequence of theoretical first k cuts $\mathbf{d}_k^\star(\mathbf{X}, \Theta)$ such that

$$d_\infty(\mathbf{d}_k^\star(\mathbf{X}, \Theta), \hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta)) \leq \xi. \tag{5.37}$$

This implies that, with probability $1-\rho$, for all n large enough and all $1 \leq j \leq k$, the j-th empirical cut $\hat{d}_{j,n}(\mathbf{X}, \Theta)$ is performed along the same coordinate as $d_j^\star(\mathbf{X}, \Theta)$.

Now, for any cell A, since the regression function is not constant on A, one can find a theoretical cut $d_A^\star$ on A such that $L^\star(d_A^\star) > 0$. Thus, the cut $d_A^\star$ is made along an informative variable, in the sense that it is performed along one of the first S variables. Consequently, for all $\mathbf{X}, \Theta$ and for all $1 \leq j \leq k$, each theoretical cut $d_j^\star(\mathbf{X}, \Theta)$ is made along one of the first S coordinates. The proof is then a consequence of inequality (5.37). □

Chapter 6

Kernel bilinear regression for toxicogenetics

Abstract We propose a new model to predict the response of human cell lines exposed to various chemicals, based on molecular characterizations of the cell's genome and transcriptome. We demonstrate the relevance of the method on the recent DREAM8 Toxicogenetics challenge.

Contents

6.1 Introduction

The response to drugs and environmental factors varies between individuals. Understanding the genetic basis for this variability and being able to identify individuals prone to adverse side effects will play an increasingly important role with the development of personalized medicines, and could contribute to the definition of appropriate regulatory limits for environmental health protection. *Toxicogenetics* is the name of the field that aims at understanding the genetic basis for individual differences in response to potential toxicants. It builds on the fast progress in our ability to genotype individuals and more generally to measure a multitude of potential biomarkers, such as mutations or gene expression, which may be useful predictors of response.

Recent efforts to systematically characterize the molecular portraits of individuals and to assess their response to various chemicals have started to generate rich collections of data, paving the way to systematic analysis in order to decipher the molecular basis for response variability and to develop predictive models of individual response. For example, a consortium of teams from the University of North Carolina (UNC), the National Institutes of Environmental Health Sciences (NIEHS), and the National Center for Advancing Translational Sciences (NCATS)

have recently generated a large population-scale toxicity screen, testing more than 150 drugs and environmental chemicals on almost a thousand cell lines derived from individuals with well-characterized genotype and transcriptome (see more details below). These data were recently released as part of a challenge whose goal was to derive a model to predict the response of new individuals to the various chemicals.

In this paper we present a machine-learning based model to solve this toxicogenetics challenge. The model implements a kernel bilinear regression, providing a principled way to integrate heterogeneous data such as genotype and transcriptome, and to leverage informations across the different chemicals tested, in a computationally efficient framework. In the rest of this paper we provide a brief description of this model and study its empirical performance on the data from the DREAM8 Toxicogenetics challenge, where it was among the best performers.

6.2 The kernel bilinear regression model

Let $\mathcal{X}$ and $\mathcal{Y}$ denote abstract vector space to represent, respectively, cell lines and chemicals. For example, if each cell line is characterized by a measure of d genetic markers, then we may take $\mathcal{X} = \mathbb{R}^d$ to represent each cell line as a vector of markers, but to keep generality we will simply assume that $\mathcal{X}$ and $\mathcal{Y}$ are endowed with positive definite kernels, respectively K_X and K_Y. Given a set of n cell lines $x_1, \ldots, x_n \in \mathcal{X}$ and p chemicals $y_1, \ldots, y_p \in \mathcal{Y}$, we assume that a quantitative measure of toxicity response $z_{i,j} \in \mathbb{R}$ has been measured when cell line x_i is exposed to chemical y_j, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Our goal is to estimate, from this data, a function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predict the response $h(x, y)$ if a cell line x is exposed to a chemical y.

We propose to model the response with a simple bilinear regression model of the form:

$$Z = f(X, Y) + b(Y) + \epsilon\,, \tag{6.1}$$

where f is a bilinear function, b is a chemical-specific bias term and ϵ is some Gaussian noise. We add the chemical-specific bias term to adjust for the large differences in absolute toxicity response values between chemicals, while the bilinear term $f(X, Y)$ can capture some patterns of variations between cell lines shared by different chemicals. We will only focus on the problem of predicting the action of known and tested chemicals on new cell lines, meaning that we will not try to estimate $b(Y)$ on new cell lines.

If x and y are finite dimensional vectors, then the bilinear term $f(x, y)$ has the simple form $x^\top M y$ for some matrix M, with Frobenius norm $\|M\|^2 = Tr(M^\top M)$. The natural generalization of this bilinear model to possibly infinite-dimensional spaces $\mathcal{X}$ and $\mathcal{Y}$ is to consider a function f in the product reproducing kernel Hilbert space $\mathcal{H}$ associated to the product kernel $K_X K_Y$, with Hilbert-Schmitt norm $\|f\|^2$. To estimate model (6.1), we solve a standard ridge regression problem:

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}^p} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(f(x_i, y_j) + b_j - z_{ij}\right)^2 + \lambda \|f\|^2\,, \tag{6.2}$$

where λ is a regularization parameter to be optimized. As shown in the next theorem, (6.2) has an analytical solution. Note that $\mathbf{1}_n$ refers to the n-dimensional vector of ones, $Diag(u)$ for

a vector $u \in \mathbb{R}^n$ refers to the $n \times n$ diagonal matrix whose diagonal is u, and $A \circ B$ for two matrices of the same size refers to their Hadamard (or entrywise) product.

Theorem 6.1. *Let $Z \in \mathbb{R}^{n \times p}$ be the response matrix, and $K_X \in \mathbb{R}^{n \times n}$ and $K_Y \in \mathbb{R}^{p \times p}$ be the kernel Gram matrices of the n cell lines and p chemicals, with respective eigenvalue decompositions $K_X = U_X D_X U_X^\top$ and $K_Y = U_Y D_Y U_Y^\top$. Let $\gamma = U_X^\top \mathbf{1}_n$ and $S \in \mathbb{R}^{n \times p}$ be defined by $S_{ij} = 1/\left(\lambda + D_X^i D_Y^j\right)$, where D_X^i (resp. D_Y^i) denotes the i-th diagonal term of D_X (resp. D_Y). Then the solution (f^*, b^*) of (6.2) is given by*

$$b^* = U_Y Diag\left(S^\top \gamma^{\circ 2}\right)^{-1} \left(S^\top \circ \left(U_Y^\top Z^\top U_X\right)\right) \gamma \tag{6.3}$$

and

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}\,, \quad f^*(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_{i,j}^* K_X(x_i, x) K_Y(y_i, y)\,, \tag{6.4}$$

where

$$\alpha^* = U_X \left(S \circ \left(U_X^\top \left(Z - \mathbf{1}_n b^{*\top}\right) U_Y\right)\right) U_Y^\top\,. \tag{6.5}$$

The most computationally expensive part to compute b^* and α^* from (6.3) and (6.5) is the eigenvalue decomposition of K_X and K_Y, and we only need to manipulate matrices of size smaller than $n \times n$ or $p \times p$. The computational complexity of the method is therefore $O(max(n, p)^3)$, and the memory requirement $O(max(n, p)^2)$.

Proof. By the representer theorem, we know that there exists a matrix $\alpha \in \mathbb{R}^{n \times p}$ such that the solution $f \in \mathcal{H}$ of (6.2) can be expanded as:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}\,, \quad f(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_{i,j} K_X(x_i, x) K_Y(y_i, y)\,. \tag{6.6}$$

Plugging back (6.4) into (6.2) and using the fact that $\| f \|^2 = Tr(\alpha^\top K_X \alpha K_Y)$ leads to the problem:

$$\min_{\alpha \in \mathbb{R}^{n \times p}, b \in \mathbb{R}^p} Tr\left((K_X \alpha K_Y + \mathbf{1}_n b^\top - Z)^\top (K_X \alpha K_Y + \mathbf{1}_n b^\top - Z) + \lambda \alpha^\top K_X \alpha K_Y\right)\,, \tag{6.7}$$

where $Tr(M)$ is the trace of the square matrix M. This is a convex, quadratic program in α and b, so we can solve it by setting its gradient to zero. The gradient in α and b are respectively:

$$\partial_\alpha = 2 K_X \left(K_X \alpha K_Y + \lambda \alpha + \mathbf{1}_n b^\top - Z\right) K_Y\,, \tag{6.8}$$

$$\partial_b = 2\left(nb + K_Y \alpha^\top K_X \mathbf{1}_n - Z^\top \mathbf{1}_n\right)\,. \tag{6.9}$$

The gradient in α (6.8) is null if and only if

$$K_X \alpha K_Y + \lambda \alpha + \mathbf{1}_n b^\top - Z = \epsilon\,, \quad \text{with} \quad K_X \epsilon K_Y = 0\,. \tag{6.10}$$

Note that although different α may satisfy (6.10) (with different ϵ), they all define the same function f through (6.6), since the original problem (6.2) is strictly convex in f and has therefore

a unique solution. We can therefore only focus on the solution corresponding to $\epsilon = 0$ in (6.10), leading to

$$K_X \alpha K_Y + \lambda \alpha = Z - \mathbf{1}_n b^\top . \tag{6.11}$$

Multiplying on the left by $U_X^\top$ and on the right by U_Y leads to

$$D_X U_X^\top \alpha U_Y D_Y + \lambda U_X^\top \alpha U_Y = U_X^\top \left(Z - \mathbf{1}_n b^\top \right) U_Y .$$

The (i,j)-th entry of the l.h.s. matrix is $\left(D_X^i D_Y^j + \lambda \right) \left(U_X^\top \alpha U_Y \right)_{i,j}$, so by denoting $S_{i,j} = \left(D_X^i D_Y^j + \lambda \right)^{-1}$ as in Theorem 6.1 we get:

$$U_X^\top \alpha U_Y = S \circ \left(U_X^\top \left(Z - \mathbf{1}_n b^\top \right) U_Y \right) ,$$

leading to

$$\alpha = U_X \left(S \circ \left(U_X^\top \left(Z - \mathbf{1}_n b^\top \right) U_Y \right) \right) U_Y^\top . \tag{6.12}$$

The gradient in b (6.9) is null if and only if

$$nb + K_Y \alpha^\top K_X \mathbf{1}_n - Z^\top \mathbf{1}_n = 0 . \tag{6.13}$$

However, (6.11) leads to

$$K_Y \alpha^\top K_X \mathbf{1}_n + \lambda \alpha^\top \mathbf{1}_n = Z^\top \mathbf{1}_n - b \mathbf{1}_n^\top \mathbf{1}_n = Z^\top \mathbf{1}_n - nb \tag{6.14}$$

which combined with (6.13) gives $\alpha^\top \mathbf{1}_n = 0$. Applying this condition to (6.12) gives

$$U_Y \left(S^\top \circ \left(U_Y^\top \left(Z^\top - b \mathbf{1}_n^\top \right) U_X \right) \right) U_X^\top \mathbf{1}_n = 0 .$$

Multiplying on the left by $U_Y^\top$ and using the notation $\gamma = U_X^\top \mathbf{1}_n$, this is equivalent to

$$\left(S^\top \circ \left(U_Y^\top Z^\top U_X \right) \right) \gamma = \left(S^\top \circ \left(U_Y^\top b \gamma^\top \right) \right) \gamma \tag{6.15}$$

Now, if we denote $c = U_Y^\top b$, then we see that the i-th entry of the vector $\left(S^\top \circ \left(c \gamma^\top \right) \right) \gamma$ is:

$$\left[\left(S^\top \circ \left(c \gamma^\top \right) \right) \gamma \right]_i = \sum_{j=1}^n S_{ji} (c_i \gamma_j) \gamma_j = c_i \sum_{j=1}^n S_{ji} \gamma_j^2 = c_i \left[S^\top \gamma^{\circ 2} \right]_i ,$$

meaning

$$\left(S^\top \circ \left(c \gamma^\top \right) \right) \gamma = Diag \left(S^\top \gamma^{\circ 2} \right) c . \tag{6.16}$$

Plugging (6.16) into (6.15) we get

$$Diag \left(S^\top \gamma^{\circ 2} \right) U_Y^\top b = \left(S^\top \circ \left(U_Y^\top Z^\top U_X \right) \right) \gamma ,$$

which finally gives

$$b = U_Y Diag \left(S^\top \gamma^{\circ 2} \right)^{-1} \left(S^\top \circ \left(U_Y^\top Z^\top U_X \right) \right) \gamma .$$

□

6.3 Data

We test our model on the data of the DREAM 8 Toxicogenetics challenge[1], a joint crowdsourcing initiative of Sage Bionetworks, DREAM and scientists at UNC, NIEHS and NCATS to assess the possibility of developing predictive models in toxicogenetics. The effect of 156 chemical compounds was tested on 884 lymphoblastoid cell lines derived from participants in the 1000 Genomes Project and representing 9 distinct geographic subpopulations. The toxicity response was measured in terms of EC10, *i.e.*, the concentration in chemical compound at which the intracellular ATP content is decreased by 10 percent. Participants had access to a subset of these data, corresponding to 106 chemicals tested 487 cell lines. The challenge we focus on was to predict the effect of these 106 chemicals on the 397 cell lines that were not given to the participants.

For each cell line we have access to three covariates (population, batch and sex), to DNA variation profiles (approximately 1.3 million single nucleotide polymorphisms or SNPs), and to gene expression levels by RNA sequencing for a subset of the cell lines (46 256 transcripts). Each chemical also came with a set of structural attributes obtained by standard chemoinformatics methodologies, including 160 descriptors based on the Chemistry Development Kit (CDK) and 9272 descriptors based on the Simplex Representation of Molecular Structure (SIRMS). In addition, we computed for each chemical 881 binary descriptors encoding the presence or absence of 881 substructures defined in the PubChem database Chen et al. [2009], and 1554 descriptors describing the ability of the chemicals to interact with 1554 human proteins known to be potential targets for drugs and xenobiotics Yamanishi et al. [2011].

To use these various, heterogeneous data in our kernel bilinear regression model, we transformed them into positive definite kernels for cell lines and chemicals. For each vector representation, we computed a linear kernel, 10 Gaussian RBF kernels with different bandwidths, and an average Gaussian RBF kernel; for the discrete cell lines covariates, we just computed linear kernels; we also added a standard graph kernel based on walks on the 2D structure of molecules [Mahé et al., 2005]. In addition, we tested various multitask kernels to leverage information across the chemicals [Evgeniou et al., 2005]: we made 11 kernels by interpolating linearly between the Dirac kernel (amounting to performing linear regression independently for the different chemicals) and the constant kernel (amounting to learning a single model for all chemicals). We also tried an empirical kernel equal to the empirical correlation matrix between the toxicity level of the different chemicals on the training cell lines. Finally, for both cell lines and chemicals, we defined an *integrated kernel* as the average of all other kernels. In total, we created 29 cell line kernels and 48 chemical kernels

6.4 Results

We first test the 29×48 combinations of cell line and chemical kernels by 5-fold cross-validation (over the cell lines) repeated 10 times on the 106 chemicals and 487 cell lines available during the challenge. Note that only 192 cell lines out of 487 had RNA-seq information; for cell lines missing RNA-seq information, we replaced any RNA-seq-based kernel by the Dirac kernel. We

[1] `https://www.synapse.org/#!Synapse:syn1761567`

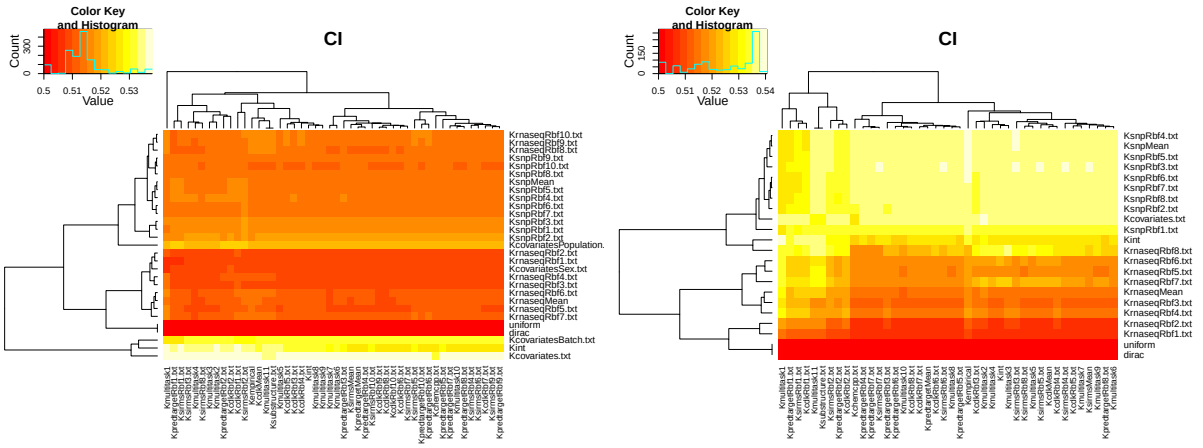

Figure 6.1: Mean CI for each combination of cell line kernel (vertical axis) and chemical kernel (horizontal axis), by cross-validation over the full set of 487 cell lines (left) or on the limited set of 192 cell lines with RNA-seq information (right).

assess the performance of prediction in terms of concordance index (CI) per chemical, averaged over the chemicals. A random prediction leads to a CI of 0.5, while a CI of 1 means that we have perfectly ranked the cell lines in terms of response to the chemical.

Figure 6.1 gives an overview of the performance reached by different combinations of kernels, with or without taking into account cell lines with missing RNA-seq data, while Figure 6.2 summarizes the average performance of each individual kernel. A first, disappointing observation is that the overall performance barely reaches $CI = 0.54$, an observation shared by participants and organizers of the challenge: this is a difficult problem. This being said, we observe clear differences between the performance of different kernels, particularly for cell lines kernels. The best performance is reached by the kernel based on covariates, in particular the batch information. This suggests that a batch effect was present in the data, but raises questions on the capacity of this kernel to generalize well to new cell lines. After the batch effect, we notice the relatively good performance of population information, captured either by the "population" covariate, or kernels based on SNP. The performance of RNA-seq was disappointing. Interestingly, the integrated kernel gave the best results besides those based on batch effects, suggesting it may be a robust and powerful way to integrated various informations. On the chemical side, we saw much less influence of kernels. We were disappointed by the fact that none of the chemical descriptors seemed to bring any performance improvement, and by the fact that the Dirac kernel (referred to as *Kmultitask1* in the plots), corresponding to fitting regression models independently for each chemical, gave the best results. However, when we examined experiments where the batch effect was less pronounced (such as the right panel of Figure 6.1), we noticed that the empirical kernel seemed to give good performance in many cases, raising hopes that the problem could benefit from some multitask strategy.

This was confirmed at the real DREAM8 Toxicogenetics challenge, where we had to predict the toxicity response of 397 new cell lines exposed to the 106 chemicals. We submitted 4 predictions with different kernels, among which the combination of integrated cell line kernel

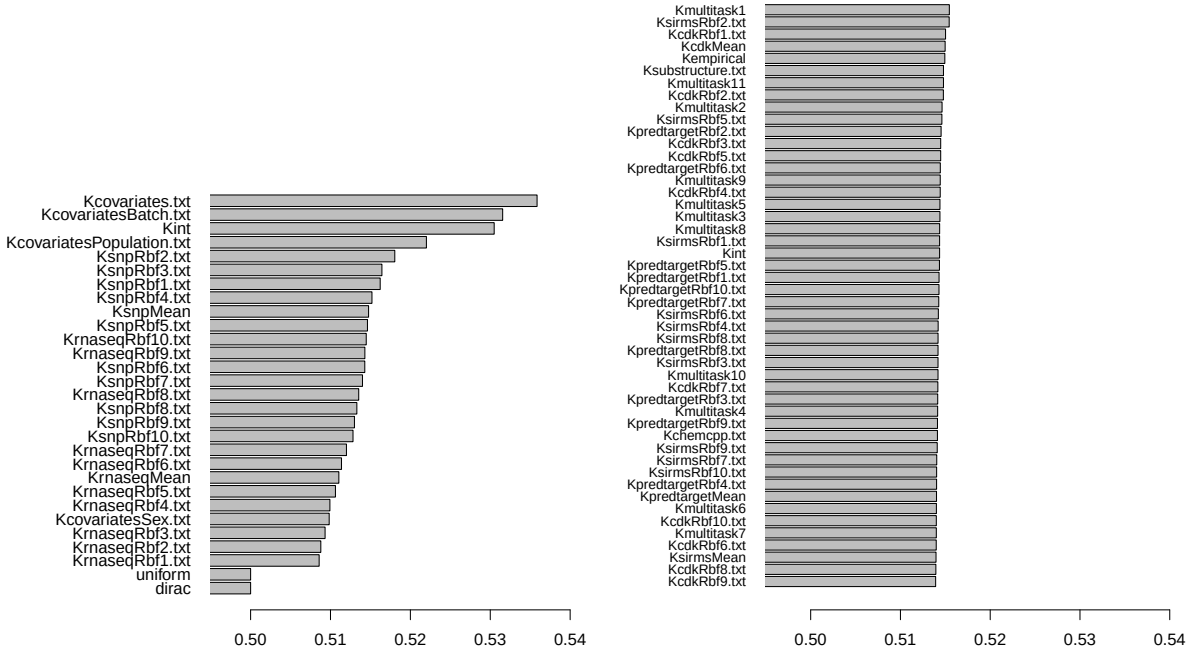

Figure 6.2: Average CI reached by each cell line kernel (left) and each chemical kernel (right).

with empirical chemical kernel performed best, with an overall rank of 2nd out of 100 submitted models.

Conclusions and perspectives

The aim of this thesis was to provide some theoretical results on random forests, in order to narrow the gap between theory and practice. Clearly random forests used in practice are built with a finite number of trees, whereas an infinite number of trees composed the forests analyzed in theory. In **Chapter 3**, we have made explicit the link between finite and infinite forest estimates by establishing their limiting distribution and proving that their errors are similar. This work is closely related to that of Mentch and Hooker [2014a] who exhibited the limiting distribution of finite random forests estimate, where both the number of trees and the number of observations tend to infinity. Wager [2014] also proved that the variance of random forests can be estimated via the Jackkniffe estimate, therefore allowing to build confidence intervals for random forest predictions. All in all, these results support the fact that theory on infinite forests can be straightforwardly extended to finite forests.

In the second part of **Chapter 3**, we studied quantile forests to highlight the benefits of random forests compared to individual trees. It is known that random forests reduce the estimation error (and, in some cases, the approximation error) of individual trees. In this context, we demonstrated that aggregating inconsistent quantile trees leads to a consistent forest. The analysis shows that quantile forest consistency results from an appropriate subsampling strategy. Since quantile forests are relatively simple to study compared to Breiman's forests, a starting point for further research could be to derive rate of convergence for quantile forests to show:

(*i*) whether quantile forests reach minimax rate of convergence for some class of regression functions;

(*ii*) whether fully developed quantile forests exhibit better performance than pruned quantile forests (no theoretical results shows the benefit of fully developed random forests).

In **Chapter 4**, we highlighted the fact that random forests can be seen as kernel estimates, up to a small modification of the algorithm. The corresponding kernels exhibit similar efficiency as their random forest counterparts while being much more interpretable. Kernels were made explicit for two purely random forests. A future line of research would be to inspect the performance of methods (SVM, Gaussian processes...) that use previous kernels as input.

Unfortunately, we did not succeed in making Breiman's forest kernels explicit (since Breiman's forest construction depends on the whole data set) and finding a closed form for Breiman's kernel does not seem to be a reasonable future goal. However, we should be able to simulate Breiman's kernel and thus empirically study their properties, particularly how they

depend on the positions $\mathbf{X}_i$ and on the labels Y_i, how they are related to purely random forest kernels (described in **Chapter 4**), and how they are affected by a particular resampling strategy (bootstrap and subsampling).

We proved in **Chapter 5** the consistency of pruned and unpruned Breiman forests. Removing assumption **(H5.2)** in **Theorem 5.2** would be a substantial improvement, although it would require to overcome technical difficulties related to the structure of CART partitioning. Nonetheless, there are other feasible avenues for further research regarding Breiman's forests:

(*i*) A complete analysis of the CART splitting criterion remains to be done. In particular, it would be very interesting to determine how the concentration of the empirical splits (near the optimal split) depends on the space dimension or the model noise, therefore leading to a better understanding of the splitting procedure.

(*ii*) Our proofs do not extend to the bootstrap resampling case whereas it is the original resampling scheme proposed by Breiman. It turns out that bootstrap distribution is not the same as the initial data distribution therefore making the original procedure far more complicated to investigate. We should empirically compute the weights $W_{ni}(\mathbf{x})$ in Breiman's forests that use bootstrap, to see if they are uniformly distributed among the Layered Nearest Neighbor of the query point $\mathbf{x}$. If this is the case, theory developed by Biau and Devroye [2010] would support the fact that random forests are consistent, even when bootstrap is applied.

(*iii*) The fact that **Theorem 5.1** and **5.2** do not provide the rate of consistency for Breiman's forests is disappointing. However our proof cannot be straightforwardly adapted to that aim. Upper bounded the rate of consistency could be possible in some particular regression model. As a consequence, the bound dependency on the ambient dimension p would help to understand how the random forest procedures are influenced by the space dimension.

Bibliography

D. Amaratunga, J. Cabrera, and Y.-S. Lee. Enriched random forests. *Bioinformatics*, 24: 2010–2014, 2008.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

K.J. Archer and R.V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52:2249–2260, 2008.

S. Arlot and R. Genuer. Analysis of purely random forests bias. arXiv:1407.3939, 2014.

L. Auret and C. Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105:157–170, 2011.

Z.-H. Bai, L. Devroye, H.-K. Hwang, and T.-H. Tsai. Maxima in hypercubes. *Random Structures & Algorithms*, 27:290–309, 2005.

M. Banerjee and I. W. McKeague. Confidence sets for split points in decision trees. *The Annals of Statistics*, 35:543–574, 2007.

O. Barndorff-Nielsen and M. Sobel. On the distribution of the number of admissible points in a vector random sample. *Theory of Probability and Its Applications*, 11:249–269, 1966.

S. Bernard, L. Heutte, and S. Adam. Forest-RK: A new random forest induction method. In D.-S. Huang, D.C. Wunsch II, D.S. Levine, and K.-H. Jo, editors, *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, pages 430–437, Berlin, 2008. Springer.

S. Bernard, S. Adam, and L. Heutte. Dynamic random forests. *Pattern Recognition Letters*, 33: 1580–1586, 2012.

G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13: 1063–1095, 2012.

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101:2499–2518, 2010.

G. Biau and L. Devroye. Cellular tree classifiers. *Electronic Journal of Statistics*, 7:1875–1912, 2013.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11:687–712, 2010.

E. G. Bongiorno, A. Goia, E. Salinelli, and P. Vieu. *Contributions in infinite-dimensional statistics and related topics*. Società Editrice Esculapio, 2014.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2:493–507, 2012.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman. *Some infinity theory for predictor ensembles*. Technical Report 577, University of California, Berkeley, 2000a.

L. Breiman. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40:229–242, 2000b.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman. *Setting up, using, and understanding random forests V4.0*. `https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf`, 2003.

L. Breiman. *Consistency for a simple model of random forests*. Technical Report 670, University of California, Berkeley, 2004.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, 1984.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30:927–961, 2002.

B. Chen, D. Wild, and R. Guha. Pubchem as a source of polypharmacology. *Journal of chemical information and modeling*, 49:2044–2055, 2009.

S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55:4316–4336, 2009.

S. Clémençon, M. Depecker, and N. Vayatis. Ranking forests. *Journal of Machine Learning Research*, 14:39–73, 2013.

A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7:81–227, 2011.

N.L. Crookston and A.O. Finley. yaImpute: An R package for kNN imputation. *Journal of Statistical Software*, 23:1–16, 2008.

A. Cutler and G. Zhao. Pert - perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497, 2001.

D.R. Cutler, T.C. Edwards Jr, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, and J.J. Lawler. Random forests for classification in ecology. *Ecology*, 88:2783–2792, 2007.

A. Davies and Z. Ghahramani. The random forest kernel and other kernels for big data from random partitions. arXiv:1402.4293, 2014.

H. Deng and G. Runger. Feature selection via regularized trees. In *The 2012 International Joint Conference on Neural Networks*, pages 1–8, 2012.

H. Deng and G. Runger. Gene selection with guided regularized random forest. *Pattern Recognition*, 46:3483–3489, 2013.

M. Denil, D. Matheson, and N. de Freitas. *Consistency of online random forests*, 2013. arXiv:1302.4853.

C. Désir, S. Bernard, C. Petitjean, and L. Heutte. One class random forests. *Pattern Recognition*, 46:3490–3506, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:1–13, 2006.

T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40:139–157, 2000a.

T. G. Dietterich. Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer, 2000b.

T. G. Dietterich and E. B. Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University, 1995.

B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38. CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, 1982.

T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. In *Journal of Machine Learning Research*, volume 6, pages 615–637, 2005.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24:543–562, 2012.

R. Genuer, J.-M. Poggi, and C. Tuleau. Random forests: some methodological insights. arXiv:0811.3619, 2008.

R. Genuer, J. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31:2225-2236, 2010.

E. Geremia, B.H. Menze, and N. Ayache. Spatially adaptive random forests. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1332–1335, 2013.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

W. Greblicki, A. Krzyzak, and M. Pawlak. Distribution-free pointwise consistency of kernel regression estimate. *The Annals of Statistics*, pages 1570–1575, 1984.

B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. arXiv:1310.5726, 2013.

B. Gregorutti, B. Michel, and P. Saint-Pierre. Grouped variable importance with random forests and application to multivariate functional data analysis. arXiv:1411.4170, 2014.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422, 2002.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

M. Hamza and D. Laroque. An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75:629–643, 2005.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–310, 1986.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Second Edition*. Springer, New York, 2009.

T. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence*, 20:832–844, 1998.

L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer, New York, 2012.

J. Howard and M. Bowles. The two most important algorithms in predictive modeling today. In *Strata Conference: Santa Clara*. `http://strataconf.com/strata2012/public/schedule/detail/22658`, 2012.

T. Ishioka. Imputation of missing values for unsupervised data using the proximity in random forests. In *eLmL 2013, The Fifth International Conference on Mobile, Hybrid, and On-line Learning*, pages 30–36. International Academy, Research, and Industry Association, 2013.

H. Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

H. Ishwaran. The effect of splitting on random forests. *Machine Learning*, pages 1–44, 2013.

H. Ishwaran and U.B. Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80:1056–1064, 2010.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2:841–860, 2008.

H. Ishwaran, U.B. Kogalur, X. Chen, and A.J. Minn. Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4:115–132, 2011.

D. Jeffrey and G. Sanja. Simplified data processing on large clusters. *Communications of the ACM*, 51:107–113, 2008.

A. Joly, P. Geurts, and L. Wehenkel. Random forests with random projections of the output space for high dimensional multi-label classification. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 607–622, Berlin, 2014. Springer.

A. Kleiner, A. Talwalkar, P. Sarkar, and M.I. Jordan. A scalable bootstrap for massive data. arXiv:1112.5016, 2012.

E. Konukoglu and M. Ganz. Approximate false positive rate control in selection frequency for random forest. *arXiv:1410.2838*, 2014.

A. Kyrillidis and A. Zouzias. Non-uniform feature sampling for decision tree ensembles. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4548–4552, 2014.

B. Lakshminarayanan, D. M. Roy, and Y. W. Teh. Mondrian forests: Efficient online random forests. arXiv:1406.2673, 2014.

P. Latinne, O. Debeir, and C. Decaestecker. Limiting the number of trees in random forests. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, pages 178–187, Berlin, 2001. Springer.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28:1302–1338, 2000.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2:18–22, 2002.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101:578–590, 2006.

G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, pages 431–439, 2013.

P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert. Graph kernels for molecular structure-activity relationship analysis with support vector machines. *Journal of chemical information and modeling*, 45:939–951, 2005.

L. Meier, S. Van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37:3779–3821, 2009.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

N. Meinshausen. Forest Garrote. *Electronic Journal of Statistics*, 3:1288–1304, 2009.

L. Mentch and G. Hooker. Ensemble trees and clts: Statistical inference for supervised learning. arXiv:1404.6473, 2014a.

L. Mentch and G. Hooker. A novel test for additivity in supervised ensemble learners. *arXiv:1406.1845*, 2014b.

E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9:141–142, 1964.

K.K. Nicodemus and J.D. Malley. Predictor correlation impacts machine learning algorithms: Implications for genomic studies. *Bioinformatics*, 25:1884–1890, 2009.

A. Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24:1084–1105, 1996.

J.-M. Poggi and C. Tuleau. Classification supervisée en grande dimension. application à l'agrément de conduite automobile. *Revue de Statistique Appliquée*, 54:41–60, 2006.

D.N. Politis, J.P. Romano, and M. Wolf. *Subsampling*. Springer, New York, 1999.

A.M. Prasad, L.R. Iverson, and A. Liaw. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9:181–199, 2006.

Y. Qi. *Ensemble Machine Learning*, chapter Random forest for bioinformatics, pages 307–323. Springer, 2012.

S.S. Qian, R.S. King, and C.J. Richardson. Two statistical methods for the detection of environmental thresholds. *Ecological Modelling*, 166:87–97, 2003.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

A. Rieger, T. Hothorn, and C. Strobl. *Random forests with missing values in the covariates*. Technical Report 79, University of Munich, Munich, 2010.

G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr. Randomized trees for human pose detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. In *IEEE 12th International Conference on Computer Vision Workshops*, pages 1393–1400, 2009.

E. Scornet. On the asymptotics of random forests. arXiv:1409.2090, 2014.

E. Scornet. Random forests and kernel methods. arXiv:1502.03836, 2015.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests: Supplementary materials. 2015a.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43:1716–1741, 2015b.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.

C.J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8:1348–1360, 1980.

C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

C.J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, pages 689–705, 1985.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008.

V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43:1947–1958, 2003.

L. Toloşi and T. Lengauer. Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, 27:1986–1994, 2011.

A. K. Y. Truong. *Fast Growing and Interpretable Oblique Trees via Logistic Regression Models.* PhD thesis, University of Oxford, 2009.

M. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes : With Applications to Statistics.* Springer, New York, 1996.

H. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28:3–28, 2014.

S. Wager. Asymptotic theory for random forests. arXiv:1405.0352, 2014.

S. Wager, T. Hastie, and B. Efron. Standard errors for bagged predictors and random forests. arXiv:1311.4555, 2013.

S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651, 2014.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

S.J. Winham, R.R. Freimuth, and J.M. Biernacka. A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6:496–505, 2013.

Y. Yamanishi, E. Pauwels, H. Saigo, and V. Stoven. Extracting sets of chemical substructures and protein domains governing drug-target interactions. *Journal of chemical information and modeling*, 51:1183–1194, 2011.

D. Yan, A. Chen, and M.I. Jordan. Cluster forests. *Computational Statistics & Data Analysis*, 66:178–192, 2013.

F. Yang, J. Wang, and G. Fan. Kernel induced random survival forests. *arXiv:1008.3952*, 2010.

Z. Yi, S. Soatto, M. Dewan, and Y. Zhan. Information forests. In *Information Theory and Applications Workshop*, pages 143–146, 2012.

R. Zhu, D. Zeng, and M.R. Kosorok. *Reinforcement learning trees.* Technical Report, University of North Carolina, Chapel Hill, 2012.

Résumé

Cette thèse est consacrée aux forêts aléatoires, une méthode d'apprentissage non paramétrique introduite par Breiman en 2001. Très répandues dans le monde des applications, les forêts aléatoires possèdent de bonnes performances et permettent de traiter efficacement de grands volumes de données. Cependant, la théorie des forêts ne permet pas d'expliquer à ce jour l'ensemble des bonnes propriétés de l'algorithme. Après avoir dressé un état de l'art des résultats théoriques existants, nous nous intéressons en premier lieu au lien entre les forêts infinies (analysées en théorie) et les forêts finies (utilisées en pratique). Nous proposons en particulier une manière de choisir le nombre d'arbres pour que les erreurs des forêts finies et infinies soient proches. D'autre part, nous étudions les forêts quantiles, un type d'algorithme proche des forêts de Breiman. Dans ce cadre, nous démontrons l'intérêt d'agréger des arbres : même si chaque arbre de la forêt quantile est inconsistant, grâce à un sous-échantillonnage adapté, la forêt quantile est consistante. Dans un deuxième temps, nous prouvons que les forêts aléatoires sont naturellement liées à des estimateurs à noyau que nous explicitons. Des bornes sur la vitesse de convergence de ces estimateurs sont également établies. Nous démontrons, dans une troisième approche, deux théorèmes sur la consistance des forêts de Breiman élaguées et complètement développées. Dans ce dernier cas, nous soulignons, comme pour les forêts quantiles, l'importance du sous-échantillonnage dans la consistance de la forêt. Enfin, nous présentons un travail indépendant portant sur l'estimation de la toxicité de certains composés chimiques.

Mots-clefs: Estimation non-paramétrique, régression, forêt aléatoire, méthodes à noyau, consistance, arbre de régression, agrégation.

Abstract

This is devoted to a nonparametric estimation method called random forests, introduced by Breiman in 2001. Extensively used in a variety of areas, random forests exhibit good empirical performance and can handle massive data sets. However, the mathematical forces driving the algorithm remain largely unknown. After reviewing theoretical literature, we focus on the link between infinite forests (theoretically analyzed) and finite forests (used in practice) aiming at narrowing the gap between theory and practice. In particular, we propose a way to select the number of trees such that the errors of finite and infinite forests are similar. On the other hand, we study quantile forests, a type of algorithms close in spirit to Breiman's forests. In this context, we prove the benefit of trees aggregation: while each tree of quantile forest is not consistent, with a proper subsampling step, the forest is. Next, we show the connection between forests and some particular kernel estimates, which can be made explicit in some cases. We also establish upper bounds on the rate of convergence for these kernel estimates. Then we demonstrate two theorems on the consistency of both pruned and unpruned Breiman forests. We stress the importance of subsampling to demonstrate the consistency of the unpruned Breiman's forests. At last, we present the results of a Dreamchallenge whose goal was to predict the toxicity of several compounds for several patients based on their genetic profile.

Keywords: random forest, kernel methods, consistency, aggregation.